SOLID PHASE SEQUENCING OF BIOPOLYMERS

RIGHTS IN THE INVENTION

This invention was made with United States Government support under grant number DE-FG02-93ER61609, awarded by the United States Department of Energy, and the United States Government has certain rights in the invention.

REFERENCE TO RELATED APPLICATIONS

This application is a continuation of United States Application Serial No. 08/420,009, filed April 11, 1995, entitled "Solid Phase Sequencing of Nucleic Acids," which is incorporated herein by reference in its entirety, including all figures, tables, and drawings. This application is also a continuation of United States Application Serial No. 08/470,835, filed June 6, 1995, entitled "Solid Phase Sequencing of Nucleic Acids," which is a continuation of United States Application Serial No. 08/420,009, each of which which is incorporated herein by reference in its entirety, including all figures, tables, and drawings. This application is also a continuation of United States Application Serial No. 08/419,994, filed April 11, 1995, entitled "Solid Phase Sequencing of Biopolymers by Mass Spectrometry," which is incorporated herein by reference in its entirety, including all figures, tables, and drawings. This application is also a continuation of United States Application Serial No. 08/470,716, filed June 6, 1995, entitled "Solid Phase Sequencing of Biopolymers by Mass Spectrometry", which is a continuation of United States Application Serial No. 08/419,994), each of which is incorporated herein by reference in its entirety, including all figures, tables, and drawings.

BACKGROUND

1. FIELD OF THE INVENTION

This invention relates to methods for detecting and sequencing nucleic acids with sequencing by hybridization technology and molecular weight analysis, to probes and probe arrays useful in sequencing and detection and to kits and apparatus for determining sequence information.

2. DESCRIPTION OF THE BACKGROUND

Since the recognition of nucleic acid as the carrier of the genetic code, a great deal of interest has centered around determining the sequence of that code in the many forms which it is found. Two landmark studies made the process of nucleic acid sequencing, at least with DNA, a common and relatively rapid procedure practiced in most laboratories. The first describes a process whereby terminally labeled DNA molecules are chemically cleaved at single base repetitions (A.M. Maxam and W. Gilbert, Proc. Natl. Acad. Sci. USA 74:560-64, 1977). Each base position in the nucleic acid sequence is then determined from the molecular weights of fragments produced by partial cleavages. Individual reactions were devised to cleave preferentially at guanine, at adenine, at cytosine and thymine, and at cytosine alone. When the products of these four reactions are resolved by molecular weight, using, for example, polyacrylamide gel electrophoresis, DNA sequences can be read from the pattern of fragments on the resolved gel.

The second study describes a procedure whereby DNA is sequenced using a variation of the plus-minus method (F. Sanger et al., Proc. Natl. Acad. Sci. USA 74:5463-67, 1977). This procedure takes advantage of the chain terminating ability of dideoxynucleoside triphosphates (ddNTPs) and the ability of DNA polymerase to incorporate ddNTP with nearly equal fidelity as the natural substrate of DNA polymerase, deoxynucleosides triphosphates (dNTPs). Briefly, a primer, usually an oligonucleotide, and a template DNA are incubated together in the presence of a useful concentration of all four dNTPs plus a limited amount of a single ddNTP. The DNA polymerase occasionally incorporates a dideoxynucleotide which terminates chain extension. Because the dideoxynucleotide has no 3'-hydroxyl, the initiation point for the polymerase enzyme is lost. Polymerization produces a mixture of fragments of varied sizes, all having identical 3' termini. Fractionation of the mixture by, for example, polyacrylamide gel electrophoresis, produces

a pattern which indicates the presence and position of each base in the nucleic acid. Reactions with each of the four ddNTPs allows one of ordinary skill to read an entire nucleic acid sequence from a resolved gel.

Despite their advantages, these procedures are cumbersome and impractical when one wishes to obtain megabases of sequence information. Further, these procedures are, for all practical purposes, limited to sequencing DNA. Although variations have developed, it is still not possible using either process to obtain sequence information directly from any other form of nucleic acid.

A relatively new method for obtaining sequence information from a nucleic acid has recently been developed whereby the sequences of groups of contiguous bases are determined simultaneously. In comparison to traditional techniques whereby one determines base-specific information of a sequence individually, this method, referred to as sequencing by hybridization (SBH), represents a many-fold amplification in speed. Due, at least in part to the increased speed, SBH presents numerous advantages including reduced expense and greater accuracy. Two general approaches of sequencing by hybridization have been suggested and their practicality has been demonstrated in pilot studies. In one format, a complete set of 4^n nucleotides of length n is immobilized as an ordered array on a solid support and an unknown DNA sequence is hybridized to this array (K.R. Khrapko et al., J. DNA Sequencing and Mapping 1:375-88, 1991). The resulting hybridization pattern provides all "*n*- tuple" words in the sequence. This is sufficient to determine short sequences except for simple tandem repeats.

In the second format, an array of immobilized samples is hybridized with one short oligonucleotide at a time (Z. Strezoska et al., Proc. Natl. Acad. Sci. USA 88:10, 089-93, 1991). When repeated $4n$ times for each oligonucleotide of length n, much of the sequence of all the immobilized samples would be determined. In both approaches, the intrinsic power of

the method is that many sequenced regions are determined in parallel. In actual practice the array size is about 10^4 to 10^5.

Another aspect of the method is that information obtained is quite redundant, and especially as the size of the nucleic acid probe grows. Mathematical simulations have shown that the method is quite resistant to experimental errors and that far fewer than all probes are necessary to determine reliable sequence data (P.A. Pevzner et al., J. Biomol. Struc. & Dyn. 9:399-410, 1991; W. Bains, Genomics 11:295-301, 1991).

In spite of an overall optimistic outlook, there are still a number of potentially severe drawbacks to actual implementation of sequencing by hybridization. First and foremost among these is that 4^n rapidly becomes quite a large number if chemical synthesis of all of the oligonucleotide probes is actually contemplated. Various schemes of automating this synthesis and compressing the products into a small scale array, a sequencing chip, have been proposed.

There is also a poor level of discrimination between a correctly hybridized, perfectly matched duplexes, and end mismatches. In part, these drawbacks have been addressed at least to a small degree by the method of continuous stacking hybridization as reported by a Khrapko et al. (FEBS Lett. 256:118-22, 1989). Continuous stacking hybridization is based upon the observation that when a single-stranded oligonucleotide is hybridized adjacent to a double-stranded oligonucleotide, the two duplexes are mutually stabilized as if they are positioned side-to-side due to a stacking contact between them. The stability of the interaction decreases significantly as stacking is disrupted by nucleotide displacement, gap or terminal mismatch. Internal mismatches are presumably ignorable because their thermodynamic stability is so much less than perfect matches. Although promising, a related problem arises which is the inability to distinguish between weak, but correct duplex formation, and simple background such as non-specific adsorption of probes to the underlying support matrix.

Detection is also monochromatic wherein separate sequential positive and negative controls must be run to discriminate between a correct hybridization match, a mismatch, and background. All too often, ambiguities develop in reading sequences longer than a few hundred base pairs on account of sequence recurrences. For example, if a sequence one base shorter than the probe recurs three times in the target, the sequence position cannot be uniquely determined. The locations of these sequence ambiguities are called branch points.

Secondary structures often develop in the target nucleic acid affecting accessibility of the sequences. This could lead to blocks of sequences that are unreadable if the secondary structure is more stable than occurs on the complementary strand.

A final drawback is the possibility that certain probes will have anomalous behavior and for one reason or another, be recalcitrant to hybridization under whatever standard sets of conditions ultimately used. A simple example of this is the difficulty in finding matching conditions for probes rich in G/C content. A more complex example could be sequences with a high propensity to form triple helices. The only way to rigorously explore these possibilities is to carry out extensive hybridization studies with all possible oligonucleotides of length "*n*" under the particular format and conditions chosen. This is clearly impractical if many sets of conditions are involved.

Among the early publications which appeared discussing sequencing by hybridization, E.M. Southern (WO 89/10977), described methods whereby unknown, or target, nucleic acids are labeled, hybridized to a set of nucleotides of chosen length on a solid support, and the nucleotide sequence of the target determined, at least partially, from knowledge of the sequence of the bound fragments and the pattern of hybridization observed. Although promising, as a practical matter, this method has numerous drawbacks. Probes are entirely single-stranded and binding stability is dependent upon the size of the duplex. However,

every additional nucleotide of the probe necessarily increases the size of the array by four fold creating a dichotomy which severely restricts its plausible use. Further, there is an inability to deal with branch point ambiguities or secondary structure of the target, and hybridization conditions will have to be tailored or in some way accounted for each binding event. Attempts have been made to overcome or circumvent these problems.

R. Drmanac et al. (U.S. Patent No. 5,202,231) is directed to methods for sequencing by hybridization using sets of oligonucleotide probes with random or variable sequences. These probes, although useful, suffer from some of the same drawbacks as the methodology of Southern (1989), and like Southern, fail to recognize the advantages of stacking interactions.

K.R. Khrapko et al. (FEBS Lett. 256:118-22, 1989; and J. DNA Sequencing and Mapping 1:357-88, 1991) attempt to address some of these problems using a technique referred to as continuous stacking hybridization. With continuous stacking, conceptually, the entire sequence of a target nucleic acid can be determined. Basically, the target is hybridized to an array of probes, again single-stranded, denatured from the array, and the dissociation kinetics of denaturation analyzed to determine the target sequence. Although also promising, discrimination between matches and mis-matches (and simple background) is low and, further, as hybridization conditions are inconstant for each duplex, discrimination becomes increasingly reduced with increasing target complexity.

Another major problem with current sequencing formats is the inability to efficiently detect sequence information. In conventional procedures, individual sequences are separated by, for example, electrophoresis using capillary or slab gels. This step is slow, expensive and requires the talents of a number of highly trained individuals, and,

more importantly, is prone to error. One attempt to overcome these difficulties has been to utilize the technology of mass spectrometry.

Mass spectrometry of organic molecules was made possible by the development of instruments able to volatize large varieties of organic compounds and by the discovery that the molecular ion formed by volatization breaks down into charged fragments whose structures can be related to the intact molecule. Although the process itself is relatively straight forward, actual implementation is quite complex. Briefly, the sample molecule or analyte is volatized and the resulting vapor passed into an ion chamber where it is bombarded with electrons accelerated to a compatible energy level. Electron bombardment ionizes the molecules of the sample analyte and then directs the ions formed to a mass analyzer. The mass analyzer, with its combination of electrical and magnetic fields, separates impacting ions according to their mass/charge (m/e) ratios. From these ratios, the molecular weights of the impacting ions can be determined and the structure and molecular weight of the analyte determined. The entire process requires less than about 20 microseconds.

Attempts to apply mass spectrometry to the analysis of biomolecules such as proteins and nucleic acids have been disappointing. Mass spectrometric analysis has traditionally been limited to molecules with molecular weights of a few thousand daltons. At higher molecular weights, samples become increasingly difficult to volatize and large polar molecules generally cannot be vaporized without catastrophic consequences. The energy requirement is so significant that the molecule is destroyed or, even worse, fragmented. Mass spectra of fragmented molecules are often difficult or impossible to read. Fragment linking order, particularly useful for reconstructing a molecular structure, has been lost in the fragmentation process. Both signal to noise ratio and resolution are significantly negatively affected. In addition, and specifically with regard to biomolecular sequencing, extreme sensitivity is

necessary to detect the single base differences between biomolecular polymers to determine sequence identity.

A number of new methods have been developed based on the idea that heat, if applied with sufficient rapidity, will vaporize the sample biomolecule before decomposition has an opportunity to take place. This rapid heating technique is referred to as plasma desorption and there are many variations. For example, one method of plasma desorption involves placing a radioactive isotope such as Californium-252 on the surface of a sample analyte which forms a blob of plasma. From this plasma, a few ions of the sample molecule will emerge intact. Field desorption ionization, another form of desorption, utilizes strong electrostatic fields to literally extract ions from a substrate. In secondary ionization mass spectrometry or fast ion bombardment, an analyte surface is bombarded with electrons which encourage the release of intact ions. Fast atom bombardment involves bombarding a surface with accelerated ions which are neutralized by a charge exchange before they hit the surface. Presumably, neutralization of the charge lessens the probability of molecular destruction, but not the creation of ionic forms of the sample. In laser desorption, photons comprise the vehicle for depositing energy on the surface to volatize and ionize molecules of the sample. Each of these techniques has had some measure of success with different types of sample molecules. Recently, there have also been a variety of techniques and combinations of techniques specifically directed to the analysis of nucleic acids.

Brennan et al. used nuclide markers to identify terminal nucleotides in a DNA sequence by mass spectrometry (U.S. Patent No. 5,003,059). Stable nuclides, detectable by mass spectrometry, were placed in each of the four dideoxynucleotides used as reagents to polymerize cDNA copies of the target DNA sequence. Polymerized copies were separated electrophoretically by size and the terminal nucleotide identified by the presence of the unique label.

Fenn et al. describes a process for the production of a mass spectrum containing a multiplicity of peaks (U.S. Patent No. 5,130,538). Peak components comprised multiply charged ions formed by dispersing a solution containing an analyte into a bath gas of highly charged droplets. An electrostatic field charged the surface of the solution and dispersed the liquid into a spray referred to as an electrospray (ES) of charged droplets. This nebulization provided a high charge/mass ratio for the droplets increasing the upper limit of volatization. Detection was still limited to less than about 100,000 daltons.

Jacobson et al. utilizes mass spectrometry to analyze a DNA sequence by incorporating stable isotopes into the sequence (U.S. Patent No. 5,002,868). Incorporation required the steps of enzymatically introducing the isotope into a strand of DNA at a terminus, electrophoretically separating the strands to determine fragment size and analyzing the separated strand by mass spectrometry. Although accuracy was stated to have been increased, electrophoresis was necessary to isolate the labeled strand.

Brennan also utilized stable markers to label the terminal nucleotides in a nucleic acid sequence, but added the step of completely degrading the components of the sample prior to analysis (U.S. Patent Nos. 5,003,059 and 5,174,962). Nuclide markers, enzymatically incorporated into either dideoxynucleotides or nucleic acid primers, were electrophoretically separated. Bands were collected and subjected to combustion and passed through a mass spectrometer. Combustion converts the DNA into oxides of carbon, hydrogen, nitrogen and phosphorous, and the label into sulfur dioxide. Labeled combustion products were identified and the mass of the initial molecule reconstructed. Although fairly accurate, the process does not lend itself to large scale sequencing of biopolymers.

A recent advancement in the mass spectrometric analysis of high molecular weight molecules in biology has been the development of time

of flight mass spectrometry (TOF-MS) with matrix-assisted laser desorption ionization (MALDI). This process involves placing the sample into a matrix which contains molecules which assist in the desorption process by absorbing energy at the frequency used to desorp the sample. The theory is that volatization of the matrix molecules encourages volatization of the sample without significant destruction. Time of flight analysis utilizes the travel time or flight time of the various ionic species as an accurate indicator of molecular mass. There have been some notable successes with these techniques.

Beavis et al. proposed to measure the molecular weights of DNA fragments in mixtures prepared by either Maxam-Gilbert or Sanger sequencing techniques (U.S. Patent No. 5,288,644). Each of the different DNA fragments to be generated would have a common origin and terminate at a particular base along an unknown sequence. The separate mixtures would be analyzed by laser desorption time of flight mass spectroscopy to determine fragment molecular weights. Spectra obtained from each reaction would be compared using computer algorithms to determine the location of each of the four bases and ultimately, the sequence of the fragment.

Williams et al. utilized a combination of pulsed laser ablation, multiphoton ionization and time of flight mass spectrometry. Effective laser desorption was accomplished by ablating a frozen film of a solution containing sample molecules. When ablated, the film produces an expanding vapor plume which entrains the intact molecules for analysis by mass spectrometry.

Even more recent developments in mass spectrometry have further increased the upper limits of molecular weight detection and determination. Mass spectrograph systems with reflectors in the flight tube have effectively doubled resolution. Reflectors also compensate for errors in mass caused by the fact that the ionized/accelerated region of the instrument is not a point source, but an area of finite size wherein

ions can accelerate at any point. Spatial differences between the origination points of the particles, problematic in conventional instruments because arrival times at the detector will vary, are overcome. Particles that spend more time in the accelerating field will also spend more time in the retarding field. Therefore, all particles emerging from the reflector should be synchronous, vastly improving resolution.

Despite these advances, it is still not possible to generate coordinated spectra representing a continuous sequence. Furthermore, throughput is sufficiently slow so as to make these methods impractical for large scale analysis of sequence information.

SUMMARY OF THE INVENTION

The present invention overcomes the problems and disadvantages associated with current strategies and designs and provides methods, kits and systems for determining the sequence of target nucleic acids.

One embodiment of the invention is directed to methods for sequencing a target nucleic acid. A set of nucleic acid fragments containing a sequence which is complementary or homologous to a sequence of the target is hybridized to an array of nucleic acid probes wherein each probe comprises a double-stranded portion, a single-stranded portion and a variable sequence within the single-stranded portion, forming a target array of nucleic acids. Molecular weights for a plurality of nucleic acids of the target array are determined and the sequence of the target constructed. Nucleic acids of the target, the target sequence, the set and the probes may be DNA, RNA or PNA comprising purine, pyrimidine or modified bases. The probes may be fixed to a solid support such as a hybridization chip to facilitate automated determination of molecular weights and identification of the target sequence.

Another embodiment of the invention is directed to methods for sequencing a target nucleic acid. A set of nucleic acid fragments containing a sequence which is complementary or homologous to a

sequence of the target is hybridized to an array of nucleic acid probes forming a target array containing a plurality of nucleic acid complexes. A strand of those probes hybridized by a fragment is extended using the fragment as template. Molecular weights of a plurality of nucleic acids of the target array are determined and the sequence of the target constructed. Strands can be enzymatically extended using chain terminating and chain elongating nucleotides. The resulting nested set of nucleic acids represents the sequence of the target. In preferred embodiments, one or more elements utilized in a method for sequencing a target nucleic acid may be mass modified. For example, elements such as probes, fragments, extended strands, chain elongating nucleotides, and/or chain terminating nucleotides may comprise at least one mass-modifying functionality.

Another embodiment of the invention is directed to methods for sequencing a target nucleic acid. A set of nucleic acid fragments containing a sequence which is complementary or homologous to a sequence of the target is hybridized to an array of mass modified probes. A strand of each probe is extended using the hybridized fragments as templates and the molecular weights of a plurality of extended and mass modified primers determined. Molecular weights for the plurality of mass modified and extended nucleic acids are determined and the sequence of the target constructed. Strands can be enzymatically extended using chain terminating and chain elongating nucleotides. The resulting nested set of nucleic acids represents the sequence of the target.

Another embodiment of the invention is directed to methods for sequencing a target nucleic acid. A set of nucleic acid fragments containing a sequence which is complementary or homologous to a sequence of the target is hybridized to an array of nucleic acid probes wherein each probe comprises a double-stranded portion, a single-stranded portion and a variable sequence within the single-stranded portion. A strand of the probe is extended and mass modified using the

hybridized fragment as a template. Molecular weights for a plurality of extended and mass modified nucleic acids are determined and the sequence of the target constructed. Nucleic acids of the target, the target sequence, the set and the probes may be DNA, RNA or PNA comprising purine, pyrimidine or modified bases. The probes may be fixed to a solid support such as a hybridization chip to facilitate automated determination of molecular weights and identification of the target sequence.

Another embodiment of the invention is directed to methods for sequencing a target nucleic acid. A set of nucleic acid fragments, each containing a sequence which corresponds to a sequence of the target is hybridized to an array of nucleic acid probes. A strand of the probe is extended using the hybridized fragment as a template. Alkali cations are removed from the extended probe, for example, by ion exchange. The molecular weights of extended strands are determined and a sequence of the target can be determined.

Another embodiment of the invention is directed to methods for sequencing a target nucleic acid. A sequence of the target is cleaved into nucleic acid fragments and the fragments hybridized to an array of nucleic acid probes. Fragments are created by enzymatically or physically cleaving the target and the sequence of the fragments is homologous with or complementary to at least a portion of the target sequence. The array is attached to a solid support and the molecular weights of the hybridized fragments determined by mass spectrometry. From the molecular weights determined, nucleotide sequences of the hybridized fragments are determined and a nucleotide sequence of the target can be identified.

Another embodiment of the invention is directed to methods for sequencing a target nucleic acid. A set of nucleic acids complementary to a sequence of the target is hybridized to an array of single-stranded nucleic acid probes wherein each probe comprises a constant sequence

and a variable sequence and said variable sequence is determinable. The molecular weights of the hybridized nucleic acids are determined and the sequence of the target identified. The array comprises less than or equal to about 4^R different probes and R is the length in nucleotides of the variable sequence and may be attached to a solid support.

Another embodiment of the invention is directed to methods for detecting a target nucleic acid. A set of nucleic acids complementary to a sequence of the target is hybridized to a fixed array of nucleic acid probes forming a target array of nucleic acid probes. The molecular weights of the hybridized nucleic acids are determined and a sequence of the target can be identified. Target nucleic acids may be obtained from biological samples such as patient samples wherein detection of the target is indicative of a disorder in the patient, such as a genetic defect, a neoplasm or an infection.

Another embodiment of the invention is directed to methods for detecting a target nucleic acid. A set of nucleic acids complementary to a sequence of the target is hybridized to a fixed array of nucleic acid probes forming a target array. A plurality of nucleic acids of the target array are mass modified and their molecular weights determined. From the molecular weights determined, nucleotide sequences of the hybridized fragments are detected. Target nucleic acids may be obtained from biological samples such as patient samples wherein detection of the target is indicative of a disorder in the patient, such as a genetic defect, a neoplasm or an infection.

Another embodiment of the invention is directed to arrays of nucleic acid probes. In these arrays, each probe comprises a first strand and a second strand wherein the first strand is hybridized to the second strand forming a double-stranded portion, a single-stranded portion and a variable sequence within the single-stranded portion. The array may be attached to a solid support such as a material that facilitates volatization of nucleic acids for mass spectrometry. Arrays can be fixed to

hybridization chips containing less than or equal to about 4^R different probes wherein R is the length in nucleotides of the variable sequence. Arrays can be used in detection methods and in kits to detect nucleic acid sequences which may be indicative of a disorder and in sequencing systems such as sequencing by mass spectrometry.

Another embodiment of the invention is directed to arrays of single-stranded nucleic acid probes wherein each probe of the array comprises a constant sequence and a variable sequence which is determinable. Arrays may be attached to solid supports which comprise matrices that facilitate volatization of nucleic acids for mass spectrometry. Arrays, generated by conventional processes, may be characterized using the above methods and replicated in mass for use in nucleic acid detection and sequencing systems.

Another embodiment of the invention is directed to arrays of mass modified nucleic acid probes. In these arrays, each probe comprises a first strand and a second strand wherein the first strand is hybridized to the second strand forming a double-stranded portion, a single-stranded portion and a variable sequence within the single-stranded portion. The array may be attached to a solid support such as a material that facilitates volatization of nucleic acids for mass spectrometry. Arrays can be fixed to hybridization chips containing less than or equal to about 4^R different probes wherein R is the length in nucleotides of the variable sequence.

Another embodiment of the invention is directed to arrays of single-stranded mass modified nucleic acid probes wherein each probe of the array comprises a constant sequence and a variable sequence which may be determinable. Arrays may be attached to solid supports which comprise matrices that facilitate volatization of nucleic acids for mass spectrometry. Arrays, generated by conventional processes, may be characterized using the above methods and replicated in mass for use in nucleic acid detection and sequencing systems.

Another embodiment of the invention is directed to kits for detecting a sequence of a target nucleic acid. Kits contain arrays of nucleic acid probes fixed to a solid support wherein each probe comprises a double-stranded portion, a single-stranded portion and a variable sequence within said single-stranded portion. The kits may contain arrays of mass modified nucleic acid probes fixed to a solid support. The solid support may be, for example, coated with a matrix that facilitates volatization of nucleic acids for mass spectrometry such as an aqueous composition.

Another embodiment of the invention is directed to mass spectrometry systems for the rapid sequencing of nucleic acids. Systems comprise a mass spectrometer, a computer with appropriate software and probe arrays which can be used to capture and sort nucleic acid sequences for subsequent analysis by mass spectrometry. The probe arrays may comprise mass modified probes.

Other embodiments and advantages of the invention are set forth, in part, in the description which follows and, in part, will be obvious from this description and may be learned from the practice of the invention.

DESCRIPTION OF THE DRAWINGS

Figure 1 (A) Schematic of a mass modified nucleic acid primer; and (B) Primer mass modification moieties.

Figure 2 (A) Schematic of mass modified nucleoside triphosphate elongators and terminators; and (B) Nucleoside triphosphate mass modification moieties.

Figure 3 List of Mass Modification Moieties.

Figure 4 List of Mass Modification Moieties.

Figure 5 Cleavage site of Mwo1 indicating bi-directional sequencing.

Figure 6 Schematic of sequencing strategy after target DNA digestion by Tsp R1.

Figure 7 Calculated T_m of Matched and Mismatched Complementary DNA.

Figure 8 Replication of a master array.

Figure 9 Reaction scheme for the covalent attachment of DNA to a surface.

Figure 10 Target nucleic acid capture and ligation.

Figure 11 Ligation efficiency of matches as compared to mismatches.

Figure 12 (A) Ligation of target DNA with probe attached at 5' Terminus; and (B) Ligation of target DNA with probe attached at 3' Terminus.

Figure 13 Gel reader sequencing results from primer hybridization analysis.

Figure 14 Mass spectrometry of oligonucleotide ladder.

Figure 15 Schematic of mass modification by alkylation.

Figure 16 Mass spectrum of 17-mer target with 0, 1 or 2 mass modified moieties.

DESCRIPTION OF THE INVENTION

As embodied and broadly described herein, the present invention is directed to methods for sequencing a nucleic acid, probe arrays useful for sequencing by mass spectrometry and kits and systems which comprise these arrays.

Nucleic acid sequencing, on both a large and small scale, is critical to many aspects of medicine and biology such as, for example, in the identification, analysis or diagnosis of diseases and disorders, and in determining relationships between living organisms. Conventional sequencing techniques rely on a base-by-base identification of the sequence using electrophoresis in a semi-solid such as an agarose or polyacrylamide gel to determine sequence identity. Although attempts have been made to apply mass spectrometric analysis to these methods, the two processes are not well suited because, at least in part,

information is still being gathered in a single base format. Sequencing-by-hybridization methodology has enhanced the sequencing process and provided a more optimistic outlook for more rapid sequencing techniques, however, this methodology is no more applicable to mass spectrometry than traditional sequencing techniques.

In contrast, positional sequencing by hybridization (PSBH) with its ability to stably bind and discriminate different sequences with large or small arrays of probes is well suited to mass spectrometric analysis. Sequence information is rapidly determined in batches and with a minimum of effort. Such processes can be used for both sequencing unknown nucleic acids and for detecting known sequences whose presence may be an indicator of a disease or contamination. Additionally, these processes can be utilized to create coordinated patterns of probe arrays with known sequences. Determination of the sequence of fragments hybridized to the probes also reveals the sequence of the probe. These processes are currently not possible with conventional techniques and, further, a coordinated batch-type analysis provides a significant increase in sequencing speed and accuracy which is expected to be required for effective large scale sequencing operations.

PSBH is also well suited to nucleic acid analysis wherein sequence information is not obtained directly from hybridization. Sequence information can be learned by coupling PSBH with techniques such as mass spectrometry. Target nucleic acid sequences can be hybridized to probes or array of probes as a method of sorting nucleic acids having distinct sequences without having *a priori* knowledge of the sequences of the various hybridization events. As each probe will be represented as multiple copies, it is only necessary that hybridization has occurred to isolate distinct sequence packages. In addition, as distinct packages of sequences, they can be amplified, modified or otherwise controlled for subsequent analysis. Amplification increases the number of specific sequences which assists in any analysis requiring increased quantities of

nucleic acid while retaining sequence specificity. Modification may involve chemically altering the nucleic acid molecule to assist with later or downstream analysis.

Consequently, another important feature of the invention is the ability to simply and rapidly mass modify the sequences of interest. A mass modification is an alteration in the mass, typically measured in terms of molecular weight as daltons, of a molecule. Mass modification which increase the discrimination between at least two nucleic acids with single base differences in size or sequence can be used to facilitate sequencing using, for example, molecular weight determinations.

One embodiment of the invention is directed to a method for sequencing a target nucleic acid using mass modified nucleic acids and mass spectrometry technology. Target nucleic acids which can be sequenced include sequences of deoxyribonucleic acid (DNA) or ribonucleic acid (RNA). Such sequences may be obtained from biological, recombinant or other man-made sources, or purified from a natural source such as a patient's tissue or obtained from environmental sources. Alternate types of molecules which can be sequenced includes polyamide nucleic acid (PNA) (P.E. Nielsen et al., Sci. 254:1497-1500, 1991) or any sequence of bases joined by a chemical backbone that have the ability to base pair or hybridize with a complementary chemical structure.

The bases of DNA, RNA and PNA include purines, pyrimidines and purine and pyrimidine derivatives and modifications, which are linearly linked to a chemical backbone. Common chemical backbone structures are deoxyribose phosphate, ribose phosphate, and polyamide. The purines of both DNA and RNA are adenine (A) and guanine (G). Others that are known to exist include xanthine, hypoxanthine, 2- and 1-diaminopurine, and other more modified bases. The pyrimidines are cytosine (C), which is common to both DNA and RNA, uracil (U) found predominantly in RNA, and thymidine (T) which occurs almost exclusively in DNA. Some of the more atypical pyrimidines include methylcytosine,

hydroxymethyl-cytosine, methyluracil, hydroxylmethyluracil, dihydroxypentyluracil, and other base modifications. These bases interact in a complementary fashion to form base-pairs, such as, for example, guanine with cytosine and adenine with thymidine. This invention also encompasses situations in which there is non-traditional base pairing such as Hoogsteen base pairing which has been identified in certain tRNA molecules and postulated to exist in a triple helix.

Sequencing involves providing a nucleic acid sequence which is homologous or complementary to a sequence of the target. Sequences may be chemically synthesized using, for example, phosphoramidite chemistry or created enzymatically by incubating the target in an appropriate buffer with chain elongating nucleotides and a nucleic acid polymerase. Initiation and termination sites can be controlled with dideoxynucleotides or oligonucleotide primers, or by placing coded signals directly into the nucleic acids. The sequence created may comprise any portion of the target sequence or the entire sequence. Alternatively, sequencing may involve elongating DNA in the presence of boron derivatives of nucleotide triphosphates. Resulting double-stranded samples are treated with a 3′ exonuclease such as exonuclease III. This exonuclease stops when it encounters a boronated residue thereby creating a sequencing ladder.

Nucleic acids can also be purified, if necessary to remove substances which could be harmful (*e.g.* toxins), dangerous (*e.g.* infectious) or might interfere with the hybridization reaction or the sensitivity of that reaction (*e.g.* metals, salts, protein, lipids). Purification may involve techniques such as chemical extraction with salts, chloroform or phenol, sedimentation centrifugation, chromatography or other techniques known to those of ordinary skill in the art.

If sufficient quantities of target nucleic acid are available and the nucleic acids are sufficiently pure or can be purified so that any substances which would interfere with hybridization are removed, a

plurality of target nucleic acids may be directly hybridized to the array. Sequence information can be obtained without creating complementary or homologous copies of a target sequence.

Sequences may also be amplified, if necessary or desired, to increase the number of copies of the target sequence using, for example, polymerase chain reactions (PCR) technology or any of the amplification procedures. Amplification involves denaturation of template DNA by heating in the presence of a large molar excess of each of two or more oligonucleotide primers and four dNTPs (dGTP, dCTP, dATP, dTTP). The reaction mixture is cooled to a temperature that allows the oligonucleotide primer to anneal to target sequences, after which the annealed primers are extended with DNA polymerase. The cycle of denaturation, annealing, and DNA synthesis, the principal of PCR amplification, is repeated many times to generate large quantities of product which can be easily identified.

The major product of this exponential reaction is a segment of double-stranded DNA whose termini are defined by the 5′ termini of the oligonucleotide primers and whose length is defined by the distance between the primers. Under normal reaction conditions, the amount of polymerase becomes limiting after 25 to 30 cycles or about one million fold amplification. Further, amplification is achieved by diluting the sample 1000 fold and using it as the template for further rounds of amplification in another PCR. By this method, amplification levels of 10^9 to 10^{10} can be achieved during the course of 60 sequential cycles. This allows for the detection of a single copy of the target sequence in the presence of contaminating DNA, for example, by hybridization with a radioactive probe. With the use of sequential PCR, the practical detection limit of PCR can be as low as 10 copies of DNA per sample.

Although PCR is a reliable method for amplification of target sequences, a number of other techniques can be used such as ligase chain reaction, self-sustained sequence replication, Qβ replicase

amplification, polymerase chain reaction linked ligase chain reaction, gapped ligase chain reaction, ligase chain detection and strand displacement amplification. The principle of ligase chain reaction is based in part on the ligation of two adjacent synthetic oligonucleotide primers which uniquely hybridize to one strand of the target DNA or RNA. If the target is present, the two oligonucleotides can be covalently linked by ligase. A second pair of primers, almost entirely complementary to the first pair of primers is also provided. The template and the four primers are placed into a thermocycler with a thermostable ligase. As the temperature is raised and lowered, oligonucleotides are renatured immediately adjacent to each other on the template and ligated. The ligated product of one reaction serves as the template for a subsequent round of ligation. The presence of target is manifested as a DNA fragment with a length equal to the sum of the two adjacent oligonucleotides.

Target sequences are fragmented, if necessary, into a plurality of fragments using physical, chemical or enzymatic means to create a set of fragments of uniform or relatively uniform length. Preferably, the sequences are enzymatically cleaved using nucleases such as DNases or RNases (mung bean nuclease, micrococcal nuclease, DNase I, RNase A, RNase T1), type I or II restriction endonucleases, or other site-specific or nonspecific endonucleases. Sizes of nucleic acid fragments are between about 5 to about 1,000 nucleotides in length, preferably between about 10 to about 200 nucleotides in length, and more preferably between about 12 to about 100 nucleotides in length. Sizes in the range of about 5, 10, 12, 15, 18, 20, 24, 26, 30 and 35 are useful to perform small scale analysis of short regions of a nucleic acid target. Fragment sizes in the range of 25, 50, 75, 125, 150, 175, 200 and 250 nucleotides and larger are useful for rapidly analyzing larger target sequences.

Target sequences may also be enzymatically synthesized using, for example, a nucleic acid polymerase and a collection of chain elongating

nucleotides (NTPs, dNTPs) and limiting amounts of chain terminating (ddNTPs) nucleotides. This type of polymerization reaction can be controlled by varying the concentration of chain terminating nucleotides to create sets, for example nested sets, which span various size ranges. In a nested set, fragments will have one common terminus and one terminus which will be different between the members of the set such that the larger fragments will contain the sequences of the smaller fragments.

The set of fragments created, which may be either homologous or complementary to the target sequence, is hybridized to an array of nucleic acid probes forming a target array of nucleic acid probe/fragment complexes. An array constitutes an ordered or structured plurality of nucleic acids which may be fixed to a solid support or in liquid suspension. Hybridization of the fragments to the array allows for sorting of very large collections of nucleic acid fragments into identifiable groups. Sorting does not require *a priori* knowledge of the sequences of the probes, and can greatly facilitate analysis by, for example, mass spectrophotometric techniques.

Hybridization between complementary bases of DNA, RNA, PNA, or combinations of DNA, RNA and PNA, occurs under a wide variety of conditions such as variations in temperature, salt concentration, electrostatic strength, and buffer composition. Examples of these conditions and methods for applying them are described in *Nucleic Acid Hybridization: A Practical Approach* (B.D. Hames and S.J. Higgins, editors, IRL Press, 1985). It is preferred that hybridization takes place between about 0°C and about 70°C, for periods of from about one minute to about one hour, depending on the nature of the sequence to be hybridized and its length. However, it is recognized that hybridizations can occur in seconds or hours, depending on the conditions of the reaction. For example, typical hybridization conditions for a mixture of two 20-mers is to bring the mixture to 68°C and let it cool to room

temperature (22°C) for five minutes or at very low temperatures such as 2°C in 2 microliters. Hybridization between nucleic acids may be facilitated using buffers such as Tris-EDTA (TE), Tris-HCl and HEPES, salt solutions (*e.g.* NaCl, KCl, $CaCl_2$), other aqueous solutions, reagents and chemicals. Examples of these reagents include single-stranded binding proteins such as Rec A protein, T4 gene 32 protein, E. coli single-stranded binding protein and major or minor nucleic acid groove binding proteins. Examples of other reagents and chemicals include divalent ions, polyvalent ions and intercalating substances such as ethidium bromide, actinomycin D, psoralen and angelicin.

Optionally, hybridized target sequences may be ligated to a single strand of the probes thereby creating ligated target-probe complexes or ligated target arrays. Ligation of target nucleic acid to probe increases fidelity of hybridization and allows for incorrectly hybridized target to be easily washed from correctly hybridized target. More importantly, the addition of a ligation step allows for hybridizations to be performed under a single set of hybridization conditions. Variation of hybridization conditions due to base composition are no longer relevant as nucleic acids with high A/T or G/C content ligate with equal efficiency. Consequently, discrimination is very high between matches and mismatches, much higher than has been achieved using other methodologies wherein the effects of G/C content were only somewhat neutralized in high concentrations of quarternary or tertiary amines such as, for example, 3M tetramethyl ammonium chloride. Further, hybridization conditions such as temperatures of between about 22°C to about 37°C, salt concentrations of between about 0.05 M to about 0.5 M, and hybridization times of between about less than one hour to about 14 hours (overnight), are also suitable for ligation. Ligation reactions can be accomplished using a eukaryotic derived or a prokaryotic derived ligase such as T4 DNA or RNA ligase. Methods for use of these and other nucleic acid modifying

enzymes are described in *Current Protocols in Molecular Biology* (F.M. Ausubel et al., editors, John Wiley & Sons, 1989).

Each probe of the probe array comprises a single-stranded portion, an optional double-stranded portion and a variable sequence within the single-stranded portion. These probes may be DNA, RNA, PNA, or any combination thereof, and may be derived from natural sources or recombinant sources, or be organically synthesized. Preferably, each probe has one or more double-stranded portions which are about 4 to about 30 nucleotides in length, preferably about 5 to about 15 nucleotides and more preferably about 7 to about 12 nucleotides, and may also be identical within the various probes of the array, one or more single-stranded portions which are about 4 to 20 nucleotides in length, preferably between about 5 to about 12 nucleotides and more preferably between about 6 to about 10 nucleotides, and a variable sequence within the single-stranded portion which is about 4 to 20 nucleotides in length and preferably about 4, 5, 6, 7 or 8 nucleotides in length. Overall probe sizes may range from as small as 8 nucleotides in lengths to 100 nucleotides and above. Preferably, sizes are from about 12 to about 35 nucleotides, and more preferably, from about 12 to about 25 nucleotides in length.

Probe sequences may be partly or entirely known, determinable or completely unknown. Known sequences can be created, for example, by chemically synthesizing individual probes with a specified sequence at each region. Probes with determinable variable regions may be chemically synthesized with random sequences and the sequence information determined separately. Either or both the single-stranded and the double-stranded regions may comprise constant sequences such as, for example, when an area of the probe or hybridized nucleic acid would benefit from having a constant sequence as a point of reference in subsequent analyses.

An advantage of this type of probe is in its structure. Hybridization of the target nucleic acid is encouraged due to the favorable thermodynamic conditions, including base-stacking interactions, established by the presence of the adjacent double strandedness of the probe. Probes may be structured with terminal single-stranded regions which consist entirely or partly of variable sequences, internal single-stranded regions which contain both constant and variable regions, or combinations of these structures. Preferably, the probe has a single-stranded region at one terminus and a double-stranded region at the opposite terminus.

Fragmented target sequences, preferably, will have a distribution of terminal sequences sufficiently broad so that the nucleotide sequence of the hybridized fragments will include the entire sequence of the target nucleic acid. Consequently, the typical probe array will comprise a collection of probes with sufficient sequence diversity in the variable regions to hybridize, with complete or nearly complete discrimination, all of the target sequence or the target-derived sequences. The resulting target array will comprise the entire target sequence on strands of hybridized probes. By way of example only, if the variable portion consisted of a four nucleotide sequence (R=4) of adenine, guanine, thymine, and cytosine, the total number of possible combinations (4^R) would be 4^4 or 256 different nucleic acid probes. If the number of nucleotides in the variable sequence was five, the number of different probes within the set would be 4^5 or 1,024. In addition, it is also possible to utilize probes wherein the variable nucleotide sequence contains gapped segments, or positions along the variable sequence which will base pair with any nucleotide or at least not interfere with adjacent base pairing.

A nucleic acid strand of the target array may be extended or elongated enzymatically. Either the hybridized fragment or one or the other of the probe strands can be extended. Extension reactions can

utilize various regions of the target array as a template. For example, when fragment sequences are longer than the hybridizable portion of a probe having a 3′ single-stranded terminus, the probe will have a 3′ overhang and a 5′ overhang after hybridization of the fragment. The now internal 3′ terminus of the one strand of the probe can be used as a primer to prime an extension reaction using, for example, an appropriate nucleic acid polymerase and chain elongating nucleotides. The extended strand of the probe will contain sequence information of the entire hybridized fragment. Reaction mixtures containing dideoxynucleotides will create a set of extended strands of varying lengths and, preferably, a nested set of strands. As the fragments have been initially sorted by hybridization to the array, each probe of the array will contain sets of nucleic acids that represent each segment of the target sequence. Base sequence information can be determined from each extended probe. Compilation of the sequence information from the array, which may require computer assistance with very large arrays, will allow one to determine the sequence of the target. Depending on the structure of the probe (*e.g.* 5′ overhang, 3′ overhang, internal single-stranded region), strands of the probe or strands of hybridized nucleic acid containing target sequence can also be enzymatically amplified by, for example, single primer PCR reactions. Variations of this process may involve aspects of strand displacement amplification, Qβ replicase amplification, self-sustained sequence replication amplification and any of the various polymerase chain reaction amplification technologies.

Extended nucleic acid strands of the probe can be mass modified using a variety of techniques and methodologies. The most straight forward may be to enzymatically synthesize the extension utilizing a polymerase and nucleotide reagents, such as mass modified chain elongating and chain terminating nucleotides. Mass modified nucleotides incorporate into the growing nucleic acid chain. Mass modifications may be introduced in most sites of the macromolecule which do not interfere

with the hydrogen bonds required for base pair formation during nucleic acid hybridization. Typical modifications include modification of the heterocyclic bases, modifications of the sugar moiety (ribose or deoxyribose), and modifications of the phosphate group. Specifically, a modifying functionality, which may be a chemical moiety, is placed at or covalently coupled to the C2, N3, N7 or C8 positions of purines, or the N7 or C9 positions of deazapurines. Modifications may also be placed at the C5 or C6 positions of pyrimidines (*e.g.* Figures 1A, 1B, 2A and 2B). Examples of useful modifying groups include deuterium, F, Cl, Br, I, SiR_3, $Si(CH_3)_3$, $Si(CH_3)_2(C_2H_5)$, $Si(CH_3)(C_2H_5)_2$, $Si(C_2H_5)_3$, $(CH_2)_nCH_3$, $(CH_2)_nNR_2$, CH_2CONR_2, $(CH_2)_nOH$, CH_2F, CHF_2, and CF_3; wherein n is an integer and R is selected from the group consisting of -H, deuterium and alkyls, alkoxys and aryls of 1-6 carbon atoms, polyoxymethylene, monoalkylated polyoxymethylene, polyethylene imine, polyamide, polyester, alkylated silyl, heterooligo/polyaminoacid and polyethylene glycol (Figures 3 and 4).

Mass modifying functionalities can include $-N_3$ or -XR, wherein X is: -O-, -NH-, -NR-, -S-, -NHC(S)-, $-OCO(CH_2)_nCOO-$, $-NHCO(CH_2)_nCOO-$, $-OSO_2O-$, $-OCO(CH_2)_n-$, -NHC(S)NH-, $-OCO(CH_2)_nS-$, $-OCO(CH_2)S-$, $-NC_4O_2H_2S-$, -OPO(O-alkyl)-, or -OP(O-alkyl)-, and n is an integer from 1 to 20; and R is: -H, deuterium and alkyls, alkoxys or aryls of 1-6 carbon atoms, such as methyl, ethyl, propyl, isopropyl, t-butyl, hexyl, benzyl, benzhydral, trityl, substituted trityl, aryl, substituted aryl, polyoxymethylene, monoalkylated polyoxymethylene, polyethylene imine, polyamide, polyester, alkylated silyl, heterooligo/polyaminoacid or polyethylene glycol. These and other mass modifying functionalities which do not interfere with hybridization can be attached to a nucleic acid either alone or in combination. Preferably, combinations of different mass modifications are utilized to maximize distinctions between nucleic acids having different sequences.

Mass modifications may be major changes of molecular weight, such as occurs with coupling between a nucleic acid and a

heterooligo/polyaminoacid, or more minor such as occurs by substituting chemical moieties into the nucleic acid having molecular masses smaller than the natural moiety. Non-essential chemical groups may be eliminated or modified using, for example, an alkylating agent such as iodoacetamide. Alkylation of nucleic acids with iodoacetamide has an additional advantage that a reactive oxygen of the 3'- position of the sugar is eliminated. This provides one less site per base for alkali cations, such as sodium, to interact. Sodium, present in nearly all nucleic acids, increases the likelihood of forming satellite adduct peaks upon ionization. Adduct peaks appear at a slightly greater mass than the true molecule which would greatly reduce the accuracy of molecular weight determinations. These problems can be addressed, in part, with matrix selection in mass spectrometric analysis, but this only helps with nucleic acids of less than 20 nucleotides. Ammonium ($+NH_3$), which can substitute for the sodium cation (^+Na) during ion exchange, does not increase adduct formation. Consequently, another useful mass modification is to remove alkali cations from the entire nucleic acid. This can be accomplished by ion exchange with aqueous solutions of ammonium such as ammonium acetate, ammonium carbonate, diammonium hydrogen citrate, ammonium tartrate and combinations of these solutions. DNA dissolved in 3 M aqueous ammonium hydroxide neutralizes all the acidic functions of the molecule. As there are no protons, there is a significant reduction in fragmentation during procedures such as mass spectrometry.

Another mass modification is to utilize nucleic acids with non-ionic polar phosphate backbones (*e.g.* PNA). Such nucleotides can be generated by oligonucleoside phosphomonothioate diesters or by enzymatic synthesis using nucleic acid polymerases and alpha- (α-) thio nucleoside triphosphate and subsequent alkylation with iodoacetamide. Synthesis of such compounds is straightforward and can be performed and the products separated and isolated by, for example, analytical HPLC.

Mass modification of arrays can be performed before or after target hybridization as the modifications do not interfere with hybridized nucleic acids or with hybridization of nucleic acids. This conditioning of the array is simple to perform and easily adaptable in bulk. Probe arrays can therefore be synthesized with no special manipulations. Only after the arrays are fixed to solid supports, just in fact when it would be most convenient to perform mass modification, would probes be conditioned.

Probe strands may also be mass modified subsequent to synthesis by, for example, contacting by treating the extended strands with an alkylating agent, a thiolating agent or subjecting the nucleic acid to cation exchange. Nucleic acid which can be modified include target sequences, probe sequences and strands, extended strands of the probe and other available fragments. Probes can be mass modified on either strand prior to hybridization. Such arrays of mass modified or conditioned nucleic acids can be bound to fragments containing the target sequence with no interference to the fidelity of hybridization. Subsequent extension of either strand of the probe, for example using Sanger sequencing techniques, and using the target sequences as templates will create mass modified extended strands. The molecular weights of these strands can be determined with excellent accuracy.

Probes may be in solution, such as in wells or on the surface of a micro-tray, or attached to a solid support. Mass modification can occur while the probes are fixed to the support, prior to fixation or upon cleavage from the support which can occur concurrently with ablation when analyzed by mass spectrometry. In this regard, it can be important which strand is released from the support upon laser ablation. Preferably, in such cases, the probe is differentially attached to the support. One strand may be permanent and the other temporarily attached or, at least, selectively releasable.

Examples of solid supports which can be used include a plastic, a ceramic, a metal, a resin, a gel and a membrane. Useful types of solid

supports include plates, beads, microbeads, whiskers, combs, hybridization chips, membranes, single crystals, ceramics and self-assembling monolayers. A preferred embodiment comprises a two-dimensional or three-dimensional matrix, such as a gel or hybridization chip with multiple probe binding sites (Pevzner et al., J. Biomol. Struc. & Dyn., 9:399-410, 1991; Maskos and Southern, Nuc. Acids Res. 20:1679-84, 1992). Hybridization chips can be used to construct very large probe arrays which are subsequently hybridized with a target nucleic acid. Analysis of the hybridization pattern of the chip can assist in the identification of the target nucleotide sequence. Patterns can be manually or computer analyzed, but it is clear that positional sequencing by hybridization lends itself to computer analysis and automation. Algorithms and software have been developed for sequence reconstruction which are applicable to the methods described herein (R. Drmanac et al., J. Biomol. Struc. & Dyn. 5:1085-1102, 1991; Pevzner, J. Biomol. Struc. & Dyn. 7:63-73, 1989).

Nucleic acid probes may be attached to the solid support by covalent binding such as by conjugation with a coupling agent or by covalent or non-covalent binding such as electrostatic interactions, hydrogen bonds or antibody-antigen coupling, or by combinations thereof. Typical coupling agents include biotin/avidin, biotin/streptavidin, *Staphylococcus aureus* protein A/IgG antibody F_c fragment, and streptavidin/protein A chimeras (Sano and Cantor, Bio/Technology 9:1378-81, 1991), or derivatives or combinations of these agents. Nucleic acids may be attached to the solid support by a photocleavable bond, an electrostatic bond, a disulfide bond, a peptide bond, a diester bond or a combination of these sorts of bonds. The array may also be attached to the solid support by a selectively releasable bond such as 4,4'-dimethoxytrityl or its derivative. Derivatives which have been found to be useful include 3 or 4 [bis (4-methoxyphenyl)]methyl-benzoic acid, N-succinimidyl-3 or 4[bis-(4-methoxyphenyl)]-methyl-benzoic acid, N-

succinimidyl- 3 or 4 [bis-(4-methoxyphenyl)]-hydroxymethyl-benzoic acid, N-succinimidyl- 3 or 4 [bis-(4-methoxyphenyl)]-chloromethyl-benzoic acid, and salts of these acids.

Binding may be reversible or permanent where strong associations would be critical. In addition, probes may be attached to solid supports via spacer moieties between the probes of the array and the solid support. Useful spacers include a coupling agent, as described above for binding to other or additional coupling partners, or to render the attachment to the solid support cleavable.

Cleavable attachments may be created by attaching cleavable chemical moieties between the probes and the solid support such as an oligopeptide, oligonucleotide, oligopolyamide, oligoacrylamide, oligoethylene glycerol, alkyl chains of between about 6 to 20 carbon atoms, and combinations thereof. These moieties may be cleaved with added chemical agents, electromagnetic radiation or enzymes. Examples of attachments cleavable by enzymes include peptide bonds which can be cleaved by proteases and phosphodiester bonds which can be cleaved by nucleases. Chemical agents such as β-mercaptoethanol, dithiothreitol (DTT) and other reducing agents cleave disulfide bonds. Other agents which may be useful include oxidizing agents, hydrating agents and other selectively active compounds. Electromagnetic radiation such as ultraviolet, infrared and visible light cleave photocleavable bonds. Attachments may also be reversible such as, for example, using reversible chemical linkages of magnetic attachments. Release and reattachment can be performed using, for example, magnetic or electrical fields.

Hybridized probes can provide direct or indirect information about the hybridized sequence. Direct information may be obtained from the binding pattern of the array wherein probe sequences are known or can be determined. Indirect information requires additional analysis of a plurality of nucleic acids of the target array. For example, a specific nucleic acid sequence will have a unique or relatively unique molecular

weight depending on its size and composition. That molecular weight can be determined, for example, by chromatography (*e.g.* HPLC), nuclear magnetic resonance (NMR), high-definition gel electrophoresis, capillary electrophoresis (*e.g.* HPCE), spectroscopy or mass spectrometry. Preferably, molecular weights are determined by measuring the mass/charge ratio with mass spectrometry technology.

Mass spectrometry of biopolymers such as nucleic acids can be performed using a variety of techniques (*e.g.* U.S. Patent Nos. 4,442,354; 4,931,639; 5,002,868; 5,130,538; 5,135,870; 5,174,962). Difficulties associated with volatization of high molecular weight molecules such as DNA and RNA have been overcome, at least in part, with advances in techniques, procedures and electronic design. Further, only small quantities of sample are needed for analysis, the typical sample being a mixture of 10 or so fragments. Quantities which range from between about 0.1 femtomole to about 1.0 nanomole, preferably between about 1.0 femtomole to about 1000 femtomoles and more preferably between about 10 femtomoles to about 100 femtomoles are typically sufficient for analysis. These amounts can be easily placed onto the individual positions of a suitable surface or attached to a support.

Another of the important features of this invention is that it is unnecessary to volatize large lengths of nucleic acids to determine sequence information. Using the methods of the invention, segments of the nucleic acid target, discretely isolated into separate complexes on the target array, can be sequenced and those sequence segments collated making it unnecessary to have to volatize the entire strand at once. Techniques which can be used to volatize a nucleic acid fragment include fast atom bombardment, plasma desorption, matrix-assisted laser desorption/ionization, electrospray, photochemical release, electrical release, droplet release, resonance ionization and combinations of these techniques.

In electrohydrodynamic ionization, thermospray, aerospray and electrospray, the nucleic acid is dissolved in a solvent and injected with the help of heat, air or electricity, directly into the ionization chamber. If the method of ionization involves a light beam, particle beam or electric discharge, the sample may be attached to a surface and introduced into the ionization chamber. In such situations, a plurality of samples may be attached to a single surface or multiple surfaces and introduced simultaneously into the ionization chamber and still analyzed individually. The appropriate sector of the surface which contains the desired nucleic acid can be moved to proximate the path of an ionizing beam. After the beam is pulsed on and the surface bound molecules are ionized, a different sector of the surface is moved into the path of the beam and a second sample, with the same or different molecule, is analyzed without reloading the machine. Multiple samples may also be introduced at electrically isolated regions of a surface. Different sectors of the chip are connected to an electrical source and ionized individually. The surface to which the sample is attached may be shaped for maximum efficiency of the ionization method used. For field ionization and field desorption, a pin or sharp edge is an efficient solid support and for particle bombardment and laser ionization, a flat surface.

The goal of ionization for mass spectroscopy is to produce a whole molecule with a charge. Preferably, a matrix-assisted laser desorption/ionization (MALDI) or electrospray (ES) mass spectroscopy is used to determine molecular weight and, thus, sequence information from the target array. It will be recognized by those of ordinary skill that a variety of methods may be used which are appropriate for large molecules such as nucleic acids. Typically, a nucleic acid is dissolved in a solvent and injected into the ionization chamber using electrohydrodynamic ionization, thermospray, aerospray or electrospray. Nucleic acids may also be attached to a surface and ionized with a beam of particles or light. Particles which have successfully used include plasma (plasma

desorption), ions (fast ion bombardment) or atoms (fast atom bombardment). Ions have also been produced with the rapid application of laser energy (laser desorption) and electrical energy (field desorption).

In mass spectrometer analysis, the sample is ionized briefly by a pulse of laser beams or by an electric field induced spray. The ions are accelerated in an electric field and sent at a high velocity into the analyzer portion of the spectrometer. The speed of the accelerated ion is directly proportional to the charge (z) and inversely proportional to the mass (m) of the ion. The mass of the molecule may be deduced from the flight characteristics of its ion. For small ions, the typical detector has a magnetic field which functions to constrain the ions stream into a circular path. The radii of the paths of equally charged particles in a uniform magnetic field is directly proportional to mass. That is, a heavier particle with the same charge as a lighter particle will have a larger flight radius in a magnetic field. It is generally considered to be impractical to measure the flight characteristics of large ions such as nucleic acids in a magnetic field because the relatively high mass to charge (m/z) ratio requires a magnet of unusual size or strength. To overcome this limitation the electrospray method, for example, can consistently place multiple ions on a molecule. Multiple charges on a nucleic acid will decrease the mass to charge ratio allowing a conventional quadrupole analyzer to detect species of up to 100,000 daltons.

Nucleic acid ions generated by the matrix assisted laser desorption/ionization only have a unit charge and because of their large mass, generally require analysis by a time of flight analyzer. Time of flight analyzers are basically long tubes with a detector at one end. In the operation of a TOF analyzer, a sample is ionized briefly and accelerated down the tube. After detection, the time needed for travel down the detector tube is calculated. The mass of the ion may be calculated from the time of flight. TOF analyzers do not require a magnetic field and can detect unit charged ions with a mass of up to 100,000 daltons. For

improved resolution, the time of flight mass spectrometer may include a reflectron, a region at the end of the flight tube which negatively accelerates ions. Moving particles entering the reflectron region, which contains a field of opposite polarity to the accelerating field, are retarded to zero speed and then reverse accelerated out with the same speed but in the opposite direction. In the use of an analyzer with a reflectron, the detector is placed on the same side of the flight tube as the ion source to detect the returned ions and the effective length of the flight tube and the resolution power is effectively doubled. The calculation of mass to charge ratio from the time of flight data takes into account of the time spent in the reflectron.

Ions with the same charge to mass ratio will typically leave the ion accelerators with a range of energies because the ionization regions of a mass spectrometer is not a point source. Ions generated further away from the flight tube, spend a longer time in the accelerator field and enter the flight tube at a higher speed. Thus ions of a single species of molecule will arrive at the detector at different times. In time of flight analysis, a longer time in the flight tube in theory provide more sensitivity, but due to the different speeds of the ions, the noise (background) will also be increased. A reflectron, besides effectively doubling the effective length of the flight tube, can reduce the error and increase sensitivity by reducing the spread of detector impingement time of a single species of ions. An ion with a higher velocity will enter the reflectron at a higher velocity and stay in the reflectron region longer than a lower velocity ion. If the reflectron electrode voltages are arranged appropriately, the peak width contribution from the initial velocity distribution can be largely corrected for at the plane of the detector. The correction provided by the reflectron leads to increased mass resolution for all stable ions, those which do not dissociate in flight, in the spectrum.

While a linear field reflectron functions adequately to reduce noise and enhance sensitivity, reflectrons with more complex field strengths

offer superior correctional abilities and a number of complex reflectrons can be used. The double stage reflectron has a first region with a weaker electric field and a second region with a stronger electric field. The quadratic and the curve field reflectron have an electric field which increases as a function of the distance. These functions, as their name implies, may be a quadratic or a complex exponential function. The dual stage, quadratic, and curve field reflectrons, while more elaborate are also more accurate than the linear reflectron.

The detection of ions in a mass spectrometer is typically performed using electron detectors. To be detected, the high mass ions produced by the mass spectrometer are converted into either electrons or low mass ions at a conversion electrode These electrons or low mass ions are then used to start the electron multiplication cascade in an electron multiplier and further amplified with a fast linear amplifier. The signals from multiple analysis of a single sample are combined to improve the signal to noise ratio and the peak shapes, which also increase the accuracy of the mass determination.

This invention is also directed to the detection of multiple primary ions directly through the use of ion cyclotron resonance and Fourier analysis. This is useful for the analysis of a complete sequencing ladder immobilized on a surface. In this method, a plurality of samples are ionized at once and the ions are captured in a cell with a high magnetic field. An RF field excites the population of ions into cyclotron orbits. Because the frequencies of the orbits are a function of mass, an output signal representing the spectrum of the ion masses is obtained. This output is analyzed by a computer using Fourier analysis which reduces the combined signal to its component frequencies and thus provides a measurement of the ion masses present in the ion sample. Ion cyclotron resonance and Fourier analysis can determine the masses of all nucleic acids in a sample. The application of this method is especially useful on a sequencing ladder.

The data from mass spectrometry, either performed singly or in parallel (multiplexed), can determine the molecular mass of a nucleic acid sample. The molecular mass, combined with the known sequence of the sample, can be analyzed to determine the length of the sample. Because different bases have different molecular weight, the output of a high resolution mass spectrometer, combined with the known sequence and reaction history of the sample, will determine the sequence and length of the nucleic acid analyzed. In the mass spectroscopy of a sequencing ladder, generally the base sequence of the primers are known. From a known sequence of a certain length, the added base of a sequence one base longer can be deduced by a comparison of the mass of the two molecules. This process is continued until the complete sequence of a sequencing ladder is determined.

Another embodiment of the invention is directed to a method for detecting a target nucleic acid. As before, a set of nucleic acids complementary or homologous to a sequence of the target is hybridized to an array of nucleic acid probes. The molecular weights of the hybridized nucleic acids determined by, for example, mass spectrometry and the nucleic acid target detected by the presence of its sequence in the sample. As the object is not to obtain extensive sequence information, probe arrays may be fairly small with the critical sequences, the sequences to be detected, repeated in as many variations as possible. Variations may have greater than 95% homology to the sequence of interest, greater than 80%, greater than 70% or greater than about 60%. Variations may also have additional sequences not required or present in the target sequence to increase or decrease the degree of hybridization. Sensitivity of the array to the target sequence is increased while reducing and hopefully eliminating the number of false positives.

Target nucleic acids to be detected may be obtained from a biological sample, an archival sample, an environmental sample or another source expected to contain the target sequence. For example, samples

may be obtained from biopsies of a patient and the presence of the target sequence is indicative of the disease or disorder such as, for example, a neoplasm or an infection. Samples may also be obtained from environmental sources such as bodies of water, soil or waste sites to detect the presence and possibly identify organisms and microorganism which may be present in the sample. The presence of particular microorganisms in the sample may be indicative of a dangerous pathogen or that the normal flora is present.

Another embodiment of the invention is directed to the arrays of nucleic acid probes useful in the above-described methods and procedures. These probes comprise a first strand and a second strand wherein the first strand is hybridized to the second strand forming a double-stranded portion, a single-stranded portion and a variable sequence within the single-stranded portion. Either or both of the strands may be mass modified. The array may be attached to a solid support such as a material that facilitates volatization of nucleic acids for mass spectrometry. Typically, arrays comprise large numbers of probes such as less than or equal to about 4^R different probes and R is the length in nucleotides of the variable sequence. When utilizing arrays for large scale sequencing, larger arrays can be used whereas, arrays which are used for detection of specific sequences may be fairly small as many of the potential sequence combinations will not be necessary.

Arrays may also comprise nucleic acid probes which are entirely single-stranded and nucleic acids which are single-stranded, but possess hairpin loops which create double-stranded regions. Such structures can function in a manner similar if not identical to the partially single-stranded probes, which comprise two strands of nucleic acid, and have the additional advantage of thermodynamic energy available in the secondary structure.

Arrays may be in solution or fixed on a solid support through streptavidinbiotin interactions or other suitable coupling agents. Arrays

may also be reversibly fixed to the solid support using, for example, chemical moieties which can be cleaved with electromagnetic radiation, chemical agents and the like. The solid support may comprise materials such as matrix chemicals which assist in the volatization process for mass spectrometric analysis. Such chemicals include nicotinic acid, 3'-hydroxypicolinic acid, 2,5-dihydroxybenzoic acid, sinapinic acid, succinic acid, glycerol, urea and Tris-HCl, pH about 7.3.

Another embodiment of the invention is directed to kits for detecting a sequence of a target nucleic acid. An array of mass modified nucleic acid probes is fixed to a solid support which may comprise a matrix chemical that facilitates volatization of nucleic acids for mass spectrometry. Kits can be used to detect diseases and disorders in biological samples by detecting specific nucleic acid sequences which are indicative of the disorder. Probes may be labeled with detectable labels which only become detectable upon hybridization with a correctly matched target sequence. Detectable labels include radioisotopes, metals, luminescent or bioluminescent chemicals, fluorescent chemicals, enzymes and combinations thereof.

Another embodiment of the invention is directed to nucleic acid sequencing systems which comprise a mass spectrometer, a computer loaded with appropriate software for analysis of nucleic acids and an array of probes which can be used to capture a target nucleic acid sequence. Systems may be manual or automated as desired. The arrays may comprise mass modified probes. The U.S. Patents noted herein are specifically incorporated by reference. The following experiments are offered to illustrate embodiments of the invention, and should not be viewed as limiting the scope of the invention.

EXAMPLES

EXAMPLE 1: PREPARATION OF TARGET NUCLEIC ACID

Target nucleic acid is prepared by restriction endonuclease cleavage of cosmid DNA. The properties of type II and other restriction nucleases that cleave outside of their recognition sequences were exploited. A restriction digestion of a 10 to 50 kb DNA sample with such an enzyme produced a mixture of DNA fragments most of which have unique ends. Recognition and cleavage sites of useful enzymes are shown in Table 1.

SUBSTITUTE SHEET (RULE 26)

Table 1
Restriction Enzymes and Recognition Sites for PSBH

Mwo I	GCNNNNN↓-NNGC CGNN-NNNNNCG↑
Esi YI	CCNNNNN↓-NNGG GGNN-NNNNNCC↑
Apa BI	GCANNNNN↓-TGC CGT-NNNNNACG↑
Mnl I	CCTCN$_7$↓ GGAGN$_6$↑
TspR I	NNCAGTGNN↓ NNGTCACNN↑
Cje I	CCANNNNNN↓-GTNNNN GGTNNNNNN-CANNNN↑
Cje PI	CCANNNNN↓-NNTCNN GGTNNNNN-NNAGNN↑

One restriction enzyme, *ApaB15*, with a 6 base pair recognition site may also be used. DNA sequencing is best served by enzymes that produce average fragment lengths comparable to the lengths of DNA sequencing ladders analyzable by mass spectrometry. At present these lengths are about 100 bases or less.

BsiY I and *Mwo I* restriction endonucleases are used together to digest DNA in preparation of PSBH. Target DNA from is cleaved to completion and complexed with PSBH probes either before or after melting. The fraction of fragments with unique ends or degenerate ends depends an the complexity of the target sequence. For example, a 10 kilobase clone would yield on average 16 fragments or a total of 32 ends since each double-stranded DNA target produces two ligatable 3' ends. With 1024 possible ends, Poisson statistics (Table 2) predict that there would be 3% degeneracies. In contrast, a 40 kilobase cosmid insert would yield 64 fragments or 128 ends, of which, 12% of these would be degenerate and a 50 kilobase sample would yield 80 fragments or 160 ends. Some of these would surely be degenerate. Up to at least 100 kilobase, the larger the target the more sequence are available from each multiplex DNA sample preparation. With a 100 kilobase target, 27% of the targets would be degenerate.

Table 2
Poisson Distribution of Restriction Enzyme Sites

Target Size	*Mwo I*		*TspRI*	
(kb)	Sequencing	Assembly	Sequencing	Assembly
10	0.97	0.60	0.94	0.94
40	0.88	0.14	0.80	0.80
100	0.73	0.01	0.57	0.57

With *BsiY I* and *Mwo I*, any restriction site that yields a unique 5 base end may be captured twice and the resulting sequence data obtained will read away from the site in both directions (Figure 5). With the knowledge of three bases of overlapping sequence at the site, this sorts all sequences into 64 different categories. With 10 kilobase targets, 60% will contain fragments and, thus sequence assembly is automatic.

Two array capture methods can be used with *Mwo I* and *BsiY I*. In the first method, conventional five base capture is used. Because the two target bases adjacent to the capture site are known, they from the restriction enzyme recognition sequence, an alternative capture strategy would build the complement of these two bases into the capture sequence. Seven base capture is thermodynamically more stable, but less discriminating against mismatches.

TspR I is another commercially available restriction enzyme with properties that are very attractive for use in PSBH-mediated Sanger sequencing. The method for using *TspR I* is shown in Figure 6. *TspR I* has a five base recognition site and cuts two bases outside this site on each strand to yield nine base 3' single-stranded overhangs. These can be captured with partially duplex probes with complementary nine base overhangs. Because only four bases are not specified by enzyme recognition, *TspR I* digest results in only 256 types of cleavage sites. With human DNA the average fragment length that should result is 1370 bases. This enzyme is ideal to generate long sequence ladders and are useful to input to long thin gel sequencing where reads up to a kilobase are common. A typical human cosmid yields about 30 *TspR I* fragments or 60 ends. Given the length distribution expected, many of these could not be sequenced fully from one end. With 256 possible overhangs, Poisson statistics (Table 2) indicate that 80% adjacent fragments can be assembled with no additional labor. Thus, very long blocks of continuous DNA sequence are produced.

Three additional restriction enzymes are also useful. These are *MnI I*, *Cje I*, and *Cje PI* (Table 1). The first has a four base site with one A+T should give smaller human DNA fragments on average than *Mwo I* or *BsiY I*. The latter two have unusual interrupted five base recognition sites and might supplement *TspR I*.

Target DNA may also be prepared by tagged PCR. It is possible to add a preselected five base 3' terminal sequence to a target DNA using a

PCR primer five bases longer than the known target sequence priming site. Samples made in this way can be captured and sequenced using the PSBH approach based on the five base tag. Biotin was used to allow purification of the complementary strand prior to use as an immobilized sequencing template. Biotin may also be placed on the tag. After capture of the duplex PCR product by streptavidin-coated magnetic microbeads, the desired strand (needed to serve as a sequencing template) could be denatured from the duplex and used to contact the entire probe array. For multiplex sample preparation, a series of different five base tagged primers would be employed, ideally in a single multiplex PCR reaction. This approach also requires knowing enough target sequence for unique PCR amplification and is more useful for shotgun sequencing or comparative sequencing than for *de novo* sequencing.

EXAMPLE 2: BASIC ASPECTS OF POSITIONAL SEQUENCING BY HYBRIDIZATION

An examination of the potential advantages of stacking hybridization has been carried out by both calculations and pilot experiments. Some calculated T_m's for perfect and mismatched duplexes are shown in Figure 7. These are based on average base compositions. The calculations revealed that the binding of a second oligomer next to a preformed duplex provides an extra stability equal to about two base pairs and that mispairing seems to have a larger consequence on stacking hybridization than it does on ordinary hybridization. Other types of mispairing are less destabilizing, but these can be eliminated by requiring a ligation step. In standard SBH, a terminal mismatch is the least destabilizing event, and leads to the greatest source of ambiguity or background. For an octanucleotide complex, an average terminal mismatch leads to a 6°C lowering in T_m. For stacking hybridization, a terminal mismatch on the side away from the pre-existing duplex, is the least destabilizing event. For a pentamer, this leads to a drop in T_m of 10°C. These considerations indicate that the discrimination power of

stacking hybridization in favor of perfect duplexes are greater than ordinary SBH.

EXAMPLE 3: PREPARATION OF MODEL ARRAYS

In a single synthesis, all 1024 possible single-stranded probes with a constant 18 base stalk followed by a variable 5 base extension can be created. The 18 base extension is designed to contain two restriction enzyme cutting sites. *Hga I* generates a 5 base, 5′ overhang consisting of the variable bases N_5. *Not I* generates a 4 base, 5′ overhang at the constant end of the oligonucleotide. The synthetic 23-mer mixture hybridized with a complementary 18-mer forms a duplex which can be enzymatically extended to form all 1024, 23-mer duplexes. These are cloned by, for example, blunt end ligation, into a plasmid which lacks *Not I* sites. Colonies containing the cloned 23-base insert are selected and each clone contains one unique sequence. DNA minipreps can be cut at the constant end of the stalk, filled in with biotinylated pyrimidines and cut at the variable end of the stalk to generate the 5 base 5′ overhang. The resulting nucleic acid is fractionated by Qiagen columns (nucleic acid purification columns) to discard the high molecular weight material. The nucleic acid probe will then be attached to a streptavidin-coated surface. This procedure could easily be automated in a Beckman Biomec or equivalent chemical robot to produce many identical arrays of probes.

The initial array contains about a thousand probes. The particular sequence at any location in the array will not be known. However, the array can be used for statistical evaluation of the signal to noise ratio and the sequence discrimination for different target molecules under different hybridization conditions. Hybridization with known nucleic acid sequences allows for the identification of particular elements of the array. A sufficient set of hybridizations would train the array for any subsequent sequencing task. Arrays are partially characterized until they have the desired properties. For example, the length of the oligonucleotide duplex, the mode of its attachment to a surface and the hybridization conditions

used can all be varied using the initial set of cloned DNA probes. Once the sort of array that works best is determined, a complete and fully characterized array can be constructed by ordinary chemical synthesis.

EXAMPLE 4: PREPARATION OF SPECIFIC PROBE ARRAYS

With positional SBH, one potential trick to compensate for some variations in stability among species due to GC content variation is to provide GC rich stacking duplex adjacent AT rich overhangs and AT rich stacking duplex adjacent GC rich overhangs. Moderately dense arrays can be made using a typical x-y robot to spot the biotinylated compounds individually onto a streptavidin-coated surface. Using such robots, it is possible to make arrays of $2 \times 10_4$ samples in 100 to 400 cm^2 of nominal surface. Commercially available streptavidin-coated beads can be adhered, permanently to plastics like polystyrene, by exposing the plastic first to a brief treatment with an organic solvent like triethylamine. The resulting plastic surfaces have enormously high biotin binding capacity because of the very high surface area that results.

In certain experiments, the need for attaching oligonucleotides to surfaces may be circumvented altogether, and oligonucleotides attached to streptavidin-coated magnetic microbeads used as already done in pilot experiments. The beads can be manipulated in microtitre plates. A magnetic separator suitable for such plates can be used including the newly available compressed plates. For example, the 18 by 24 well plates (Genetix, Ltd.; USA Scientific Plastics) would allow containment of the entire array in 3 plates. This format is well handled by existing chemical robots. It is preferable to use the more compressed 36 by 48 well format so the entire array would fit on a single plate. The advantages of this approach for all the experiments are that any potential complexities from surface effects can be avoided and already-existing liquid handling, thermal control and imaging methods can be used for all the experiments.

Lastly, a rapid and highly efficient method to print arrays has been developed. Master arrays are made which direct the preparation of replicas or appropriate complementary arrays. A master array is made manually (or by a very accurate robot) by sampling a set of custom DNA sequences in the desired pattern and then transferring these sequences to the replica. The master array is just a set of all 1024-4096 compounds printed by multiple headed pipettes and compressed by offsetting. A potentially more elegant approach is shown in Figure 8. A master array is made and used to transfer components of the replicas in a sequence-specific way. The sequences to be transferred are designed to contain the desired 5 or 6 base 5′ variable overhang adjacent to a unique 15 base DNA sequence.

The master array consists of a set of streptavidin bead-impregnated plastic coated metal pins. Immobilized biotinylated DNA strands that consist of the variable 5 or 6 base segment plus the constant 15 base segment are at each tip. Any unoccupied sites on this surface are filled with excess free biotin. To produce a replica chip, the master array is incubated with the complement of the 15 base constant sequence, 5′-labeled with biotin. Next, DNA polymerase is used to synthesize the complement of the 5 or 6 base variable sequence. Then the wet pin array is touched to the streptavidin-coated surface of the replica and held at a temperature above the T_m of the complexes on the master array. If there is insufficient liquid carryover from the pin array for efficient sample transfer, the replica array could first be coated with spaced droplets of solvent, either held in concave cavities or delivered by a multiheaded pipettor. After the transfer, the replica chip is incubated with the complement of 15 base constant sequence to reform the double-stranded portions of the array. The basic advantage of this scheme is that the master array and transfer compounds are made only once and the manufacture of replica arrays can proceed almost endlessly.

EXAMPLE 5: ATTACHMENT OF NUCLEIC ACIDS PROBES TO SOLID SUPPORTS

Nucleic acids may be attached to silicon wafers or to beads. A silicone solid support was derivatized to provide iodoacetyl functionalities on its surface. Derivatized solid support were bound to disulfide containing oligodeoxynucleotides. Alternatively, the solid support may be coated with streptavidin or avidin and bound to biotinylated DNA.

Covalent attachment of oligonucleotide to derivatized chips: silicon wafers are chips with an approximate weight of 50 mg. To maintain uniform reaction condition, it was necessary to determine the exact weight of each chip and select chips of similar weights for each experiment. The reaction scheme for this procedure is shown in Figure 9.

To derivatize the chip to contain the iodoacetyl functionality an anhydrous solution of 25% (by volume) 3-aminopropyltrieshoxysilane in toluene was prepared under argon and aliquotted (700 μl) into tubes. A 50 mg chip requires approximately 700 μl of silane solution. Each chip was flamed to remove any surface contaminants during its manufacture and dropped into the silane solution. The tube containing the chip was placed under an argon environment and shaken for approximately three hours. After this time, the silane solution was removed and the chips were washed three times with toluene and three times with dimethyl sulfoxide (DMSO). A 10 mM solution of N-succinimidyl(4-iodoacetyl)aminobenzoate (SIAB) (Pierce Chemical Co.; Rockford, IL) was prepared in anhydrous DMSO and added to the tube containing a chip. Tubes were shaken under an argon environment for 20 minutes. The SIAB solution was removed and after three washes with DMSO, the chip was ready for attachment to oligonucleotides.

Some oligonucleotides were labeled so the efficiency of attachment could be monitored. Both 5' disulfide containing oligodeoxynucleotides and unmodified oligodeoxynucleotides were radiolabeled using terminal deoxynucleotidyl transferase enzyme and standard techniques. In a

typical reaction, 0.5 mM of disulfide-containing oligodeoxynucleotide mix was added to a trace amount of the same species that had been radiolabeled as described above. This mixture was incubated with dithiothreitol (DTT) (6.2 μmol, 100 mM) and ethylenediaminetetraacetic acid (EDTA) pH 8.0 (3 μmol, 50 mM). EDTA served to chelate any cobalt that remained from the radiolabeling reaction that would complicate the cleavage reaction. The reaction was allowed to proceed for 5 hours at 37°C. With the cleavage reaction essentially complete, the free thiol containing oligodeoxynucleotide was isolated using a Chromaspin-10 column.

Similarly, Tris-(2-carboxyethyl)phosphine (TCEP) (Pierce Chemical Co.; Rockford, IL) has been used to cleave the disulfide. Conditions utilize TCEP at a concentration of approximately 100 mM in pH 4.5 buffer. It is not necessary to isolate the product following the reaction since TCEP does not competitively react with the iodoacetyl functionality.

To each chip which had been derivatized to contain the iodoacetyl functionality was added to a 10 μM solution of the oligodeoxynucleotide at pH 8. The reaction was allowed to proceed overnight at room temperature. In this manner, two different oligodeoxynucleotides have been examined for their ability to bind to the iodoacetyl silicon wafer. The first was the free thiol containing oligodeoxynucleotide already described. In parallel with the free thiol containing oligodeoxynucleotide reaction, a negative control reaction has been performed that employs a 5′ unmodified oligodeoxynucleotide. This species has similarly been 3′ radiolabeled, but due to the unmodified 5′ terminus, the non-covalent, non-specific interactions may be determined. Following the reaction, the radiolabeled oligodeoxynucleotides were removed and the chips were washed 3 times with water and quantitation proceeded.

To determine the efficiency of attachment, chips of the wafer were exposed to a phosphorimager screen (Molecular Dynamics). This exposure usually proceeded overnight, but occasionally for longer periods

of time depending on the amount of radioactivity incorporated. For each different oligodeoxynucleotide utilized, reference spots were made on polystyrene in which the molar amount of oligodeoxynucleotide was known. These reference spots were also exposed to the phosphorimager screen. Upon scanning the screen, the quantity (in moles) of oligodeoxynucleotide bound to each chip was determined by comparing the counts to the specific activities of the references. Using the weight of each chip, it is possible to calculate the area of the chip:

$$(\text{g of chip})\ (1130\ \text{mm}^2/\text{g}) = x\ \text{mm}^2$$

By incorporating this value, the amount of oligodeoxynucleotide bound to each chip may be reported in fmol/mm^2. It is necessary to divide this value by two since a radioactive signal of ^{32}P is strong enough to be read through the silicon wafer. Thus the instrument is essentially recording the radioactivity from both sides of the chip.

Following the initial quantitation each chip was washed in 5 x SSC buffer (75 mM sodium citrate, 750 mM sodium chloride, pH 7) with 50% formamide at 65°C for 5 hours. Each chip was washed three times with warm water, the 5 x SSC wash was repeated, and the chips requantitated. Disulfide linked oligonucleotides were removed from the chip by incubation with 100 mM DTT at 37°C for 5 hours.

EXAMPLE 6: ATTACHMENT OF NUCLEIC ACIDS TO STREPTAVIDIN COATED SOLID SUPPORT

Immobilized single-stranded DNA targets for solid-phase DNA sequencing were prepared by PCR amplification. PCR was performed on a Perkin Elmer Cetus DNA Thermal Cycler using Vent^R (exo$^-$) DNA polymerase (New England Biolabs; Beverly, MA), and dNTP solutions (Promega; Madison, WI). EcoR I digested plasmid NB34 (a PCR™ II plasmid with a one kb anonymous human DNA insert) was used as the DNA template for amplification. PCR was performed with an 18-nucleotide upstream primer and a downstream 5'-end biotinylated 18-nucleotide primer. PCR amplification was carried out in a 100 μl or 400 μl

volume containing 10 mM KCl, 20 mM Tris-HCl (pH 8.8 at 25°C), 10 mM $(NH_4)_2SO_4$, 2 mM $MgSO_4$, 0.1% Triton X-100, 250 μM dNTPs, 2.5 μM biotinylated primer, 5μM non-biotinylated primer, less than 100 ng of plasmid DNA, and 6 units of Vent (exo$^-$) DNA polymerase per 100 μl of reaction volume. Thirty temperature cycles were performed which included a heat denaturation step at 94°C for 1 minute, followed by annealing of primers to the template DNA for 1 minute at 60°C, and DNA chain extension with Vent (exo$^-$) polymerase for 1 minute at 72°C. For amplification with the tagged primer, 45°C was selected for primer annealing. The PCR product was purified through a Ultrafree-MC 30,000 NMWL filter unit (Millipore; Bedford, MA) or by electrophoresis and extraction from a low melting agarose gel. About 10 pmol of purified PCR fragment was mixed with 1 mg of prewashed Dynabeads M280 with streptavidin (Dynal, Norway) in 100 μl of 1 M NaCl and TE incubating at 37°C or 45°C for 30 minutes. The immobilized biotinylated double-stranded DNA fragment was converted to single-stranded form by treating with freshly prepared 0.1 M NaOH at room temperature for 5 minutes. The magnetic beads, with immobilized single-stranded DNA, were washed with 0.1 M NaOH and TE.

EXAMPLE 7: HYBRIDIZATION SPECIFICITY

Hybridization was performed using probes with five and six base pair overhangs, including a five base pair match, a five base pair mismatch, a six base pair match, and a six base pair mismatch. These sequences are depicted in Table 3.

Table 3
Hybridized Test Sequences

Test Sequences:

5 bp overlap, perfect match:

3'-TCG AGA ACC TTG (SEQ ID NO 1)

GCT*-5'

3'-CTA CTA GGC TGC GTA GTC (SEQ ID NO 2)

5'-biotin-GAT GAT CCG ACG CAT CAG AGC TC-3' (SEQ ID NO 3)

5 bp overlap, mismatch at 3' end:

3'-TCG AGA ACC TTG (SEQ ID NO 1)

GCT*-5'

3'-CTA CTA GGC TGC GTA GTC (SEQ ID NO 2)

5'-biotin-GAT GAT CCG ACG CAT CAG AGC TT-3' (SEQ ID NO 4)

6 bp overlap, perfect match:

3'-TCG AGA ACC TTG (SEQ ID NO 1)

GCT*-5'

3'-CTA CTA GGC TGC GTA GTC (SEQ ID NO 2)

5'-biotin-GAT GAT CCG ACG CAT CAG AGC TCT-3' (SEQ ID NO 5)

6 bp overlap, mismatch four bases from 3' end:

3'-TCG AGA ACC TTG (SEQ ID NO 1)

GCT*-5'

3'-CTA CTA GGC TGC GTA GTC (SEQ ID NO 2)

5'-biotin-GAT GAT CCG ACG CAT CAG AGT TCT-3' (SEQ ID NO 6)

The biotinylated double-stranded probe was prepared in TE buffer by annealing the complimentary single strands together at 68°C for five minutes followed by slow cooling to room temperature. A five-fold excess of monodisperse, polystyrene-coated magnetic beads (Dynal) coated with streptavidin was added to the double-stranded probe, which as then incubated with agitation at room temperature for 30 minutes. After ligation, the samples were subjected to two cold (4°C) washes followed by one hot (90°C) wash in TE buffer (Figure 10). The ratio of ^{32}P in the hot supernatant to the total amount of ^{32}P was determined (Figure 11). At high NaCl concentrations, mismatched target sequences were either not annealed or were removed in the cold washes. Under the same conditions, the matched target sequences were annealed and ligated to the probe. The final hot wash removed the non-biotinylated probe oligonucleotide. This oligonucleotide contained the labeled target if the target had been ligated to the probe.

EXAMPLE 8: COMPENSATING FOR VARIATIONS IN BASE COMPOSITION

The dependence on T_M on base composition, and on base sequence may be overcome with the use of salts like tetramethyl ammonium halides or betaines. Alternatively, base analogs like 2,6-diamino purine and 5-bromo U can be used instead of A and T, respectively, to increase the stability of A-T base pairs, and derivatives like 7-deazaG can be used to decrease the stability of G-C base pairs. The initial Experiments shown in Table 2 indicate that the use of enzymes will eliminate many of the complications due to base sequences. This gives the approach a very significant advantage over non-enzymatic methods which require different conditions for each nucleic acid and are highly matched to GC content.

Another approach to compensate for differences in stability is to vary the base next to the stacking site. Experiments were performed to test the relative effects of all four bases in this position on overall hybridization discrimination and also on relative ligation discrimination. Other base analogs such as dU (deoxyuridine) and 7-deazaG may also be used to suppress effects of secondary structure.

EXAMPLE 9: DNA LIGATION TO OLIGONUCLEOTIDE ARRAYS.

E. coli and T4 DNA ligases can be used to covalently attach hybridized target nucleic acid to the correct immobilized oligonucleotide probe. This is a highly accurate and efficient process. Because ligase absolutely requires a correctly base paired 3′ terminus, ligase will read only the 3′-terminal sequence of the target nucleic acid. After ligation, the resulting duplex will be 23 base pairs long and it will be possible to remove unhybridized, unligated target nucleic acid using fairly stringent washing conditions. Appropriately chosen positive and negative controls demonstrate the specificity of this method, such as arrays which are lacking a 5′-terminal phosphate adjacent to the 3′ overhang since these probes will not ligate to the target nucleic acid.

There are a number of advantages to a ligation step. Physical specificity is supplanted by enzymatic specificity. Focusing on the 3' end of the target nucleic also minimizes problems arising from stable secondary structures in the target DNA. DNA ligases are also used to covalently attach hybridized target DNA to the correct immobilized oligonucleotide probe. Several tests of the feasibility of the ligation method shown in Figure 12. Biotinylated probes were attached at 5' ends (Figure 12A) or 3' ends (Figure 12B) to streptavidin-coated magnetic microbeads, and annealed with a shorter, complementary, constant sequence to produce duplexes with 5 or 6 base single-stranded overhangs. ^{32}P-end labeled targets were allowed to hybridize to the probes. Free targets were removed by capturing the beads with a magnetic separator. DNA ligase was added and ligation was allowed to proceed at various salt concentrations. The samples were washed at room temperature, again manipulating the immobilized compounds with a magnetic separator to remove non-ligated material. Finally, samples were incubated at a temperature above the T_m of the duplexes, and eluted single-strand was retained after the remainder of the samples were removed by magnetic separation. The eluate at this point consisted of the ligated material. The fraction of ligation was estimated as the amount of ^{32}P recovered in the high temperature wash versus the amount recovered in both the high and low temperature washes. Results indicated that salt conditions can be found where the ligation proceeds efficiently with perfectly matched 5 or 6 base overhangs, but not with G-T mismatches. The results of a more extensive set of similar experiments are shown in Tables 4-6.

Table 4 looks at the effect of the position of the mismatch and Table 5 examines the effect of base composition on the relative discrimination of perfect matches verses weakly destabilizing mismatches. These data demonstrate that effective discrimination between perfect matches and single mismatches occurs with all five base overhangs

tested and that there is little if any effect of base composition on the amount of ligation seen or the effectiveness of match/mismatch discrimination. Thus, the serious problems of dealing with base composition effects on stability seen in ordinary SBH do not appear to be a problem for positional SBH. Furthermore, as the worst mismatch position was the one distal from the phosphodiester bond formed in the ligation reaction, any mismatches that survived in this position would be eliminated by a polymerase extension reaction. A polymerase such as Sequenase version 2, that has no 3'- endonuclease activity or terminal transferase activity would be useful in this regard. Gel electrophoresis analysis confirmed that the putative ligation products seen in these tests were indeed the actual products synthesized.

Table 4
Ligation Efficiency of Matched and Mismatched Duplexes in 0.2 M NaCl at 37°C

(SEQ ID NO 1) 3'-TCG AGA ACC TTG GCT-5'

		Ligation Efficiency	
	CTA CTA GGC TGC GTA GTC-5'		(SEQ ID NO 2)
5'-B-	GAT GAT CCG ACG CAT CAG AGC TC	0.170	(SEQ ID NO 3)
5'-B-	GAT GAT CCG ACG CAT CAG AGC TT	0.006	(SEQ ID NO 4)
5'-B-	GAT GAT CCG ACG CAT CAG AGC TA	0.006	(SEQ ID NO 7)
5'-B-	GAT GAT CCG ACG CAT CAG AGC CC	0.002	(SEQ ID NO 8)
5'-B-	GAT GAT CCG ACG CAT CAG AGT TC	0.004	(SEQ ID NO 9)
5'-B-	GAT GAT CCG ACG CAT CAG AAC TC	0.001	(SEQ ID NO 10)

Table 5
Ligation Efficiency of Matched and Mismatched Duplexes in 0.2M NaCl at 37°C and its Dependence on AT Content of the Overhang

Overhang Sequences		AT Content	Ligation Efficiency
Match	GGCCC	0/5	0.30
Mismatch	GGCCT		0.03
Match	AGCCC	1/5	0.36
Mismatch	AGCTC		0.02
Match	AGCTC	2/5	0.17
Mismatch	AGCTT		0.01
Match	AGATC	3/5	0.24
Mismatch	AGATT		0.01
Match	ATATC	4/5	0.17
Mismatch	ATATT		0.01
Match	ATATT	5/5	0.31
Mismatch	ATATC		0.02

Table 6
Increasing Discrimination by Sequencing Extension at 37°C

Probe†		Ligation Efficiency (fraction)	Ligation Extension (cpm) (+)	(-)
CTA CTA GGC TGC GTA GTC-5' 5'-B-GAT GAT CCG ACG CAT CAG AGA TC	(SEQ ID NO 2) (SEQ ID NO 11)	0.24	4,934	29,500
CTA CTA GGC TGC GTA GTC-5' 5'-B-GAT GAT CCG ACG CAT CAG AGC TT	(SEQ ID NO 2) (SEQ ID NO 4)	0.01	116	250
Discrimination		x24	x42	x118
CTA CTA GGC TGC GTA GTC-5' 5'-B-GAT GAT CCG ACG CAT CAG ATA TC	(SEQ ID NO 2) (SEQ ID NO 12)	0.17	12,250	25,200
CTA CTA GGC TGC GTA GTC-5' 5'-B-GAT GAT CCG ACG CAT CAG ATA TT	(SEQ ID NO 2) (SEQ ID NO 13)	0.01	240	390
Discrimination		x17	x51	x65

† = target nucleic acid is hybridized to probe and has the following sequence:
3'-TCG AGA ACC TTG GCT-5'* (SEQ ID NO 1)

B = Biotin

* = radioactive label

The discrimination for the correct sequence is not as great with an external mismatch (which would be the most difficult case to discriminate) as with an internal mismatch (Table 6). A mismatch right at the ligation point would presumably offer the highest possible discrimination. In any event, the results shown are very promising. Already there is a level of discrimination with only 5 or 6 bases of overlap that is better than the discrimination seen in conventional SBH with 8 base overlaps.

EXAMPLE 10: CAPTURE AND SEQUENCING OF A TARGET NUCLEIC ACID

A mixture of target DNA was prepared by mixing equal molar ratio of eight different oligos. For each sequencing reaction, one specific partially duplex probe and eight different targets were used. The sequence of the probe and the targets are shown in Tables 7 and 8.

Table 7
Duplex Probes Used

(DF25)	5′-F-GATGATCCGACGCATCAGCTGTG 3′-CTACTAGGCTGCGTAGTC	(SEQ ID NO 14) (SEQ ID NO 2)
(DF37)	5′-F-GATGATCCGACGCATCACTCAAC 3′-CTACTAGGCTGCGTAGTG	(SEQ ID NO 15) (SEQ ID NO 2)
(DF22)	5′-F-GATGATCCGACGCATCAGAATGT 3′-CTACTAGGCTGCGTAGTC	(SEQ ID NO 16) (SEQ ID NO 2)
(DF28)	5′-F-GATGATCCGACGCATCAGCCTAG 3′-CTACTAGGCTGCGTAGTC	(SEQ ID NO 17) (SEQ ID NO 2)
(DF36)	5′-F-GATGATCCGACGCATCAGTCGAC 3′-CTACTAGGCTGCGTAGTC	(SEQ ID NO 18) (SEQ ID NO 2)
(DF11a)	5′-F-GATGATCCGACGCATCACAGCTC 3′-CTACTAGGCTGCGTAGTG	(SEQ ID NO 19) (SEQ ID NO 2)
(DF8a)	5′-F-GATGATCCGACGCATCAAGGCCC 3′-CTACTAGGCTGCGTAGTT	(SEQ ID NO 20) (SEQ ID NO 2)

Table 8
Mixture of Targets

Match		
(NB4)	3′-TTACACCGGATCGAGCCGGGTCGATCTAG (DF22)	(SEQ ID NO 21)
(NB4.5)	3′-GGATCGACCGGGTCGATCTAG (DF28)	(SEQ ID NO 22)
(DF5)	3′-AGCTGCCGGATCGAGCCGGGTCGATCTAG (DF36)	(SEQ ID NO 23)
(TS10)	3′-TCGAGAACCTTGGCT (DF11a)	(SEQ ID NO 24)
(NB3.10)	3′-CCGGGTCGATCTAG (DF8a)	(SEQ ID NO 25)

Match / Mismatch			
(NB3.4)	3'-CCGGATCAAGCCGGGTCGATCTAG	(DF8a)	(SEQ ID NO 26)
(NB3.7)	3'-TCAAGCCGGGTCGATCTAG	(DF11a)	(SEQ ID NO 27)
(NB3.9)	3'-AGCCGGGTCGATCTAG	(DF36)	(SEQ ID NO 28)

Two pmol of each of the two duplex-probe forming oligonucleotides and 1.5 pmol of each of the eight different targets were mixed in a 10 μl volume containing 2 μl of Sequenase buffer stock (200 mM Tris-HCl, pH 7.5, 100 mM $MgC1_2$, and 250 mM NaCl) from the Sequenase kit. The annealing mixture was heated to 65°C and allowed to cool slowly to room temperature. While the reaction mixture was kept on ice, 1 μl 0.1 M dithiothreitol solution, 1 μl Mn buffer (0.15 M sodium isocitrate and 0.1 M $MnCl_2$), and 2 μl of diluted Sequenase (1.5 units) were mixed, and the 2 μl of reaction mixture was added to each of the four termination mixes at room temperature (each consisting of 3 μl of the appropriate termination mix: 16 μM dATP, 16 μM dCTP, 16 μM dGTP, 16 μM dTTP and 3.2 μM of one of the four ddNTPs, in 50 mM NaCl). The reaction mixtures were further incubated at room temperature for 5 minutes, and terminated with the addition of 4 μl of Pharmacia stop mix (deionized formamide containing dextran blue 6 mg/ml). Samples were denatured at 90-95°C for 3 minutes and stored on ice prior to loading. Sequencing samples were analyzed on an ALF DNA sequencer (Pharmacia Biotech; Piscataway, NJ) using a 10% polyacrylamide gel containing 7 M urea and 0.6 x TBE. Sequencing results from the gel reader are shown in Figure 13 and summarized in Table 9. Matched targets hybridized correctly and are sequenced, whereas mismatched targets do not hybridize and are not sequenced.

Table 9
Summary of Hybridization Data

Reaction	Hybridization	Sequence	Comment
1	Probe: DF25 Target: mixture	No	mismatch
2	Probe: DF37 Target: mixture	No	mismatch
3	Probe: DF22 Target: mixture	Yes	match
4	Probe: DF28 Target: mixture	Yes	match
5	Probe: DF36 Target: mixture	Yes	match
6	Probe: DF11a Target: mixture	Yes	match
7	Probe: DF8a Target: mixture	Yes	match
8	Probe: DF8a Target: NB3.4	No	mismatch
9	Probe: DF8a Target: TS12	No	mismatch
10	Probe: DF37 Target: DF5	No	mismatch

EXAMPLE 11: ELONGATION OF NUCLEIC ACIDS BOUND TO SOLID SUPPORTS

Elongation was carried out either by using Sequenase version 2.0 kit or an AutoRead sequencing kit (Pharmacia Biotech; Piscataway, NJ) employing T7 DNA polymerase. Elongation of the immobilized single-stranded DNA target was performed with reagents from the sequencing kits for Sequenase Version 2.0 or T7 DNA polymerase. A duplex DNA probe containing a 5-base 3' overhang was used as a primer. The duplex has a 5'- fluoroscein labeled 23-mer, containing an 18-base 5' constant region and a 5-base 3' variable region (which has the same sequence as the 5'-end of the corresponding nonbiotinylated primer for PCR amplification of target DNA, and an 18-mer complementary to the constant region of the 23-mer. The duplex was formed by annealing 20 pmol of each of the two oligonucleotides in a 10 μl volume containing 2 μl of Sequenase buffer stock (200 mM Tris-HCl, pH 7.5, 100 mM $MgCl_2$, and 250 mM NaCl) from the Sequenase kit or in a 13 μl volume containing 2 μl of the annealing buffer (1 M Tris-HCl, pH 7.6, 100 mM $MgCl_2$) from the AutoRead sequencing kit. The annealing mixture was heated to 65°C and allowed to cool slowly to 37°C over a 20-30 minute

time period. The duplex primer was annealed with the immobilized single-stranded DNA target by adding the annealing mixture to the DNA-containing magnetic beads and the resulting mixture was further incubated at 37°C for 5 minutes, room temperature for 10 minutes, and finally 0°C for at least 5 minutes. For Sequenase reactions, 1 μl 0.1 M dithiothreitol solution, 1 μl Mn buffer (0.15 M sodium isocitrate and 0.1 M $MnCl_2$) for the relative short target, and 2 μl of diluted Sequenase (1.5 units) were added, and the reaction mixture was divided into four ice cold termination mixes (each consists of 3 μl of the appropriate termination mix: 80 μM dATP, 80 μM dCTP, 80 μM dGTP, 80 μM dTTP and 8 μM of one of the four ddNTPs, in 50 mM NaCl). For T7 DNA polymerase reactions, 1 μl of extension buffer (40 mM $McCl_2$, pH 7.5, 304 mM citric acid and 324 mM DTT) and 1 μl of T7 DNA polymerase (8 units) were mixed, and the reaction volume was split into four ice cold termination mixes (each consisting of 1 μl DMSO and 3 μl of the appropriate termination mix: 1 mM dATP, 1 mM dCTP, 1 mM dGTP, 1 mM dTTP and 5 μM of one of the four ddNTPs, in 50 mM NaCl and 40 mM Tris-HCl, pH 7.4). The reaction mixtures for both enzymes were further incubated at 0°C for 5 minutes, room temperature for 5 minutes and 37°C for 5 minutes. After the completion of extension, the supernatant was removed, and the magnetic beads were re-susupended in 10 μl of Pharmacia stop mix. Samples were denatured at 90-95°C for 5 minutes (under this harsh condition, both DNA template and the dideoxy fragments are released from the beads) and stored on ice prior to loading. A control experiment was performed in parallel using a 18-mer complementary to the 3' end of target DNA as the sequencing primer instead of the duplex probe, and the annealing of 18-mer to its target was carried out in a similar way as the annealing of the duplex probe.

EXAMPLE 12: CHAIN ELONGATION OF TARGET SEQUENCES

Sequencing of immobilized target DNA can be performed with Sequenase Version 2.0. A total of 5 elongation reactions, one with each of 4 dideoxy nucleotides and one with all four simultaneously, are performed. A sequencing solution, containing (40 mM Tris-HCl, pH 7.5, 20 mM $MgCl_2$, and 50 mM NaCl, 10 mM dithiothreitol solution, 15 mM sodium isocitrate and 10 mM $MnCl_2$, and 100 u/ml of Sequenase (1.5 units) is added to the hybridized target DNA. dATP, dCTP, dGTP and dTTP are added to 20 μM to initiate the elongation reaction. In the separate reactions, one of four ddNTP is added to reach a concentration of 8 μM. In the combined reaction all four ddNTP are added to the reaction to 8 μM each. The reaction mixtures were incubated at 0°C for 5 minutes, room temperature for 5 minutes, and 37°C for 5 minutes. After the completion of extension, the supernatant was removed and the elongated DNA washed with 2 mM EDTA to terminate elongation reactions. Reaction products are analyzed by mass spectrometry.

EXAMPLE 13: CAPILLARY ELECTROPHORETIC ANALYSIS OF TARGET NUCLEIC ACID

Molecular weights of target sequences may also be determined by capillary electrophoresis. A single base capillary electrophoresis instrument can be used to monitor the performance of sample preparations in high performance capillary electrophoresis sequencing. This instrument is designed so that it is easily converted to multiple channel (wavelengths) detection.

An individual element of the sample array may be engineered directly to serve as the sample input to a capillary. Typical capillaries are 250 microns o.d. and 75 microns i.d. The sample is heated or denatured to release the DNA ladder into a liquid droplet. The silicon array surfaces is ideal for this purpose. The capillary can be brought into contact with the droplet to load the sample.

To facilitate loading of large numbers of samples simultaneously or sequentially, there are two basic methods. With 250 micron o.d. capillaries it is feasible to match the dimensions of the target array and the capillary array. Then the two could be brought into contact manually or even by a robot arm using a jig to assure accurate alignment. An electrode may be engineered directly into each sector of the silicon surface so that sample loading would only require contact between the surface and the capillary array.

The second method is based on an inexpensive collection system to capture fractions eluted from high performance capillary electrophoresis. Dilution is avoided by using designs which allow sample collection without a perpendicular sheath flow. The same apparatus designed as a sample collector can also serve inversely as a sample loader. In this case, each row of the sample array, equipped with electrodes, is used directly to load samples automatically on a row of capillaries. Using either method, sequence information is determined and the target sequence constructed.

EXAMPLE 14: MASS SPECTROMETRY OF NUCLEIC ACIDS

Nucleic acids to be analyzed by mass spectrometry were redissolved in ultrapure water (MilliQ, Millipore) using amounts to obtain a concentration of 10 pmoles/μl as stock solution. An aliquot (1 μl) of this concentration or a dilution in ultrapure water was mixed with 1μl of the matrix solution on a flat metal surface serving as the probe tip and dried with a fan using cold air. In some experiments, cation-ion exchange beads in the acid form were added to the mixture of matrix and sample solution to stabilize ions formed during analysis.

MALDI-TOF spectra were obtained on different commercial instruments such as Vision 2000 (Finnigan-MAT), VG TofSpec (Fisons Instruments), LaserTec Research (Vestec). The conditions were linear negative ion mode with an acceleration voltage of 25 kV. Mass calibration was done externally and generally achieved by using defined

peptides of appropriate mass range such as insulin, gramicidin S, trypsinogen, bovine serum albumen and cytochrome C. All spectra were generated by employing a nitrogen laser with 5 nanosecond pulses at a wavelength of 337 nm. Laser energy varied between 10^6 and 10^7 W/cm^2. To improve signal-to-noise ratio generally, the intensities of 10 to 30 laser shots were accumulated. The output of a typical mass spectrometry showing discrimination between nucleic acids which differ by one base is shown in Figure 14.

EXAMPLE 15: SEQUENCE DETERMINATION FROM MASS SPECTROMETRY

Elongation of a target nucleic acid, in the presence of dideoxy chain terminating nucleotides, generated four families of chain-terminated fragments. The mass difference per nucleotide addition is 289.19 for dpC, 313.21 for dpA, 329.21 for dpG and 304.20 for dpT, respectively. Comparison of the mass differences measured between fragments with the known masses of each nucleotide the nucleic acid sequence can be determined. Nucleic acid may also be sequenced by performing polymerase chain elongation in four separate reactions each with one dideoxy chain terminating nucleotide. To examine mass differences, 13 oligonucleotides from 7 to 50 bases in length were analyzed by MALDI-TOF mass spectrometry. The correlation of calculated molecular weights of the ddT fragments of a Sanger sequencing reaction and their experimentally verified weights are shown in Table 10. When the mass spectrometry data from all four chain termination reactions are combined, the molecular weight difference between two adjacent peaks can be used to determine the sequence.

Table 10
Summary of Molecular Weights Expected v. Measured

Fragment (n-mer)	Calculated Mass	Experimental Mass	Difference
7-mer	2104.45	2119.9	+15.4
10-mer	3011.04	3026.1	+15.1
11-mer	3315.24	3330.1	+14.9
19-mer	5771.82	5788.0	+16.2
20-mer	6076.02	6093.8	+17.8
24-mer	7311.82	7374.9	+63.1
26-mer	7945.22	7960.9	+15.7
33-mer	10112.63	10125.3	+12.7
37-mer	11348.43	11361.4	+13.0
38-mer	11652.62	11670.2	+17.6
42-mer	12872.42	12888.3	+15.9
46-mer	14108.22	14125.0	+16.8
50-mer	15344.02	15362.6	+18.6

SUBSTITUTE SHEET (RULE 26)

EXAMPLE 16: REDUCED PASS SEQUENCING

To maximize the use of PSBH arrays to produce Sanger ladders, the sequence of a target should be covered as completely as possible with the lowest amount of initial sequencing redundancy. This will maximize the performance of individual elements of the arrays and maximize the amount of useful sequence data obtained each time an array is used. With an unknown DNA, a full array of 1024 elements (*Mwo I* or *BsiY I* cleavage) or 256 elements (*TspR I* cleavage) is used. A 50 kb target DNA is cut into about 64 fragments by *Mwo I* or *BsiY I* or 30 fragments by *TspR I*, respectively. Each fragment has two ends both of which can be captured independently. The coverage of each array after capture and ignoring degeneracies is 128/1024 sites in the first case and 60/256 sites in the second case. Direct use of such an array to blindly deliver samples

element by element for mass spectrometry sequencing would be inefficient since most array elements will have no samples.

In one method, phosphatased double-stranded targets are used at high concentrations to saturate each array element that detects a sample. The target is ligated to make the capture irreversible. Next a different sample mixture is exposed to the array and subsequently ligated in place. This process is repeated four or five times until most of the elements of the array contain a unique sample. Any tandem target-target complexes will be removed by a subsequent ligating step because all of the targets are phosphatased.

Alternatively, the array may be monitored by confocal microscopy after the elongation reactions. This should reveal which elements contain elongated nucleic acids and this information is communicated to an automated robotic system that is ultimately used to load the samples onto a mass spectrometry analyzer.

EXAMPLE 17: SYNTHESIS OF MASS MODIFIED NUCLEIC ACID PRIMERS

Mass modification at the 5' sugar

Oligonucleotides were synthesized by standard automated DNA synthesis using β-cyanoethylphosphoamidites and a 5'-amino group introduced at the end of solid phase DNA synthesis. The total amount of an oligonucleotide synthesis, starting with 0.25 μmol CPG-bound nucleoside, is deprotected with concentrated aqueous ammonia, purified via OligoPAK™ Cartridges (Millipore; Bedford, MA) and lyophilized. This material with a 5'-terminal amino group is dissolved in 100 μl absolute N,N-dimethylformamide (DMF) and condensed with 10 μmol N-Fmoc-glycine pentafluorophenyl ester for 60 minutes at 25°C. After ethanol precipitation and centrifugation, the Fmoc group is cleaved off by a 10 minute treatment with 100 μl of a solution of 20% piperidine in N,N-dimethylformamide. Excess piperidine, DMF and the cleavage product from the Fmoc group are removed by ethanol precipitation and the

precipitate lyophilized from 10 mM TEAA buffer pH 7.2. This material is now either used as primer for the Sanger DNA sequencing reactions or one or more glycine residues (or other suitable protected amino acid active esters) are added to create a series of mass-modified primer oligonucleotides suitable for Sanger DNA or RNA sequencing.

Mass modification at the heterocyclic base with glycine

Starting material was 5-(3-aminopropynyl-I)-3'5'-di-p-tolyldeoxyuridine prepared and 3'5'-de-O-acylated (Haralambidis et al., Nuc. Acids Res. 15:4857-76, 1987). 0.281 g (1.0 mmole) 5-(3-aminopropynyl-I)-2'deoxyuridine were reacted with 0.927 g (2.0 mmole) N-Fmoc-glycine pentafluorophenylester in 5 ml absolute N,N-dimethylformamide in the presence of 0.129g (1 mmole; 174μl) N,N-diisopropylethylamine for 60 minutes at room temperature. Solvents were removed by rotary evaporation and the product was purified by silica gel chromatography (Kieselgel 60, Merck; column: 2.5 x 50 cm, elution with chloroform/methanol mixtures). Yield was 0.44 g (0.78 mmole, 78%). To add another glycine residue, the Fmoc group is removed with a 20 minutes treatment with 20% solution of piperidine in DMF, evaporated *in vacuo* and the remaining solid material extracted three times with 20 ml ethylacetate. After having removed the remaining ethylacetate, N-Fmoc-glycine pentafluorophenylester is coupled as described above. 5-(3(N-Fmoc-glycyl)-amidopropynyl-I)-2'-deoxyuridine is transformed into the 5'-O-dimethoxytritylated nucleoside-3'-O-β-cyanoethyl N,N-diisopropylphosphoamidite and incorporated into automated oligonucleotide synthesis. This glycine modified thymidine analogue building block for chemical DNA synthesis can be used to substitute one or more of the thymidine/uridine nucleotides in the nucleic acid primer sequence. The Fmoc group is removed at the end of the solid phase synthesis with a 20 minute treatment with a 20% solution of piperidine in DMF at room temperature. DMF is removed by a washing step with acetonitrile and the oligonucleotide deprotected and purified.

Mass modification at the heterocyclic base with β-alanine

0.281 g (1.0 mmole) 5-(3-Aminopropynyl-I)-2'-deoxyuridine was reacted with N-Fmoc-β-alanine pentafluorophenylester (0.955 g, 2.0 mmole) in 5 ml N,N-dimethylformamide (DMF) in the presence of 0.129 g (174 µl; 1.0 mmole) N,N-disopropylethylamine for 60 minutes at room temperature. Solvents were removed and the product purified by silica gel chromatography. Yield was 0.425 g (0.74 mmole; 74%). Another β-alanine moiety can be added in exactly the same way after removal of the Fmoc group. The preparation of the 5' O-dimethoxytritylatednucleoside-3'-O-β-cyanoethyl-N,N-diisopropylphosphoamidite from 5-(3-(N-Fmoc-β-alanyl)-amidopropynyl-I)-2'-deoxyuridine and incorporation into automated oligonucleotide synthesis is performed under standard conditions. This building block can substitute for any of the thymidine/uridine residues in the nucleic acid primer sequence.

Mass modification at the heterocyclic base with ethylene monomethyl ether

5-(3-aminopropynyl-I)-2'-deoxyuridine was used as a nucleosidic component in this example. 7.61 g (100.0 mmole) freshly distilled ethylene glycol monomethyl ether dissolved in 50 ml absolute pyridine was reacted with 10.01 g (100.0 mmole) recrystallized succinic anhydride in the presence of 1.22 g (10.0 mmole) 4-N,N-dimethylaminopyridine overnight at room temperature. The reaction was terminated by the addition of water (5.0 ml), the reaction mixture evaporated *in vacuo*, co-evaporated twice with dry toluene (20 ml each) and the residue redissolved in 100 ml dichloromethane. The solution was twice extracted successively with 10% aqueous citric acid (2 x 20 ml) and once with water (20 ml) and the organic phase dried over anhydrous sodium sulfate. The organic phase was evaporated *in vacuo*. Residue was redissolved in 50 ml dichloromethane and precipitated into 500 ml pentane and the precipitate dried *in vacuo*. Yield was 13.12 g (74.0 mmole; 74%). 8.86 g (50.0 mmole) of succinylated ethylene glycol monomethyl ether was

dissolved in 100 ml dioxane containing 5% dry pyridine (5 ml) and 6.96 g (50.0 mmole) 4-nitrophenol and 10.32 g (50.0 mmole) dicyclohexylcarbodiimide was added and the reaction run at room temperature for 4 hours. Dicyclohexylurea was removed by filtration, the filtrate evaporated *in vacuo* and the residue redissolved in 50 ml anhydrous DMF. 12.5 ml (about 12.5 mmole 4-nitrophenylester) of this solution was used to dissolve 2.81 g (10.0 mmole) 5-(3-aminopropynyl-l)-2'-deoxyuridine. The reaction was performed in the presence of 1.01 g (10.0 mmole; 1.4 ml) triethylamine overnight at room temperature. The reaction mixture was evaporated *in vacuo*, co-evaporated with toluene, redissolved in dichloromethane and chromatographed on silicagel (Si60, Merck; column 4 x 50 cm) with dichloromethane/methanol mixtures. Fractions containing the desired compound were collected, evaporated, redissolved in 25 ml dichloromethane and precipitated into 250 ml pentane. The dried precipitate of 5-(3-N-(O-succinyl ethylene glycol monomethyl ether)-amidopropynyl-l)-2'-deoxyuridine (yield 65%) is 5'-O-dimethoxytritylated and transformed into the nucleoside-3'-O-β-cyanoethyl-N, N-diisopropylphosphoamidite and incorporated as a building block in the automated oligonucleotide synthesis according to standard procedures. The mass modified nucleotide can substitute for one or more of the thymidine/uridine residues in the nucleic acid primer sequence. Deprotection and purification of the primer oligonucleotide also follows standard procedures.

Mass modification at the heterocyclic base with diethylene glycol monomethyl ether

Nucleosidic starting material was as in previous examples, 5-(3-aminopropynyl-l)-2'-deoxyuridine. 12.02g (100.0 mmole) freshly distilled diethylene glycol monomethyl ether dissolved in 50 ml absolute pyridine was reacted with 10.01 g (100.0 mmole) recrystallized succinic anhydride in the presence of 1.22 g (10.0 mmole) 4-N, N-dimethylaminopyridine (DMAP) overnight at room temperature. Yield was 18.35 g (82.3 mmole;

82.3%). 11.06 g (50.0 mmole) of succinylated diethylene glycol monomethyl ether was transformed into the 4-nitrophenylester and, subsequently, 12.5 mmole was reacted with 2.81 g (10.0 mmole) of 5-(3-aminopropynyl-l)-2'-deoxyuridine. Yield after silica gel column chromatography and precipitation into pentane was 3.34 g (6.9 mmole; 69%). After dimethoxytritylation and transformation into the nucleoside-β-cyanoethylphosphoamidite, the mass-modified building block is incorporated into automated chemical DNA synthesis. Within the sequence of the nucleic acid primer, one or more of the thymidine/uridine residues can be substituted by this mass-modified nucleotide.

Mass Modification at the heterocyclic base with glycine

Starting material was N^6-benzoyl-8-bromo-5'-O-(4,4'-dimethoxytrityl)-2'deoxyadenosine (Singh et al., Nuc. Acids Res. 18:3339-45, 1990). 632.5 mg (1.0 mmole) of this 8-bromo-deoxyadenosine derivative was suspended in 5 ml absolute ethanol and reacted with 251.2 mg (2.0 mmole) glycine methyl ester (hydrochloride) in the presence of 241.4 mg (2.1 mmole; 366 μl) N,N-diisopropylethylamine and refluxed until the starting nucleosidic material had disappeared (4-6 hours) as checked by thin layer chromatography (TLC). The solvent was evaporated and the residue purified by silica gel chromatography (column 2.5 x 50 cm) using solvent mixtures of chloroform/methanol containing 0.1% pyridine. Product fractions were combined, the solvent evaporated, the fractions dissolved in 5 ml dichloromethane and precipitated into 100 ml pentane. Yield was 487 mg (0.76 mmole; 76%). Transformation into the corresponding nucleoside β-cyanoethylphosphoamidite and integration into automated chemical DNA synthesis is performed under standard conditions. During final deprotection with aqueous concentrated ammonia, the methyl group is removed from the glycine moiety. The mass modified building block can substitute one or more deoxyadenosine/adenosine residues in the nucleic acid primer sequence.

Mass modification at the heterocyclic base with glycyl-glycine

632.5 mg (1.0 mmole) N^6-Benzoyl-8-bromo-5'-O-(4,4'-dimethoxytrityl)2'-deoxyadenosine was suspended in 5 ml absolute ethanol and reacted with 324.3 mg (2.0 mmole) glycylglycine methyl ester in the presence of 241.4 mg (2.1 mmole; 366 µl) N, N-diisopropylethylamine. The mixture was refluxed and completeness of the reaction checked by TLC. Yield after silica gel column chromatography and precipitation into pentane was 464 mg (0.65 mmole, 65%). Transformation into the nucleoside-β-cyanoethylphosphoamidite and into synthetic oligonucleotides is done according to standard procedures.

Mass Modification at the heterocyclic base with glycol monomethyl ether

Starting material was 5'-O-(4,4-dimethoxytrityl)-2'-amino-2'-deoxythymidine synthesized (Verheyden et al., J. Org. Chem. 36:250-54, 1971; Sasaki et al, J. Org. Chem. 41:3138-43, 1976; Imazawa et al., J. Org. Chem. 44:2039-41, 1979; Hobbs et al., J. Org. Chem. 42:714-19, 1976; and Ikehara et al., Chem. Pharm. Bull. Japan 26:240-44, 1978). 5'-O-(4,4-dimethoxytrityl)-2'-amino-2'-deoxythymidine (559.62 mg; 1.0 mmole) was reacted with 2.0 mmole of the 4-nitrophenyl ester of succinylated ethylene glycol monomethyl ether in 10 ml dry DMF in the presence of 1.0 mmole (140 µl) triethylamine for 18 hours at room temperature. The reaction mixture was evaporated *in vacuo*, co-evaporated with toluene, redissolved in dichloromethane and purified by silica gel chromatography (Si60, Merck; column: 2.5 x 50 cm; eluent: chloroform/methanol mixtures containing 0.1% triethylamine). The product containing fractions were combined, evaporated and precipitated into pentane. Yield was 524 mg (0.73 mmol; 73%). Transformation into the nucleoside-β-cyanoethyl-N,N-diisopropylphosphoamidite and incorporation into the automated chemical DNA synthesis protocol is performed by standard procedures. The mass-modified deoxythymidine derivative can substitute for one or more of the thymidine residues in the nucleic acid primer.

In an analogous way, by employing the 4-nitrophenyl ester of succinylated diethylene glycol monomethyl ether and triethylene glycol monomethyl ether, the corresponding mass-modified oligonucleotides are prepared. In the case of only one incorporated mass-modified nucleoside within the sequence, the mass difference between the ethylene, diethylene and triethylene glycol derivatives is 44.05, 88.1 and 132.15 daltons, respectively.

Mass modification at the heterocyclic base by alkylation

Phosphorothioate-containing oligonucleotides were prepared (Gait et al., Nuc. Acids Res. 19:1183, 1991). One, several or all internucleotide linkages can be modified in this way. The (-)M13 nucleic acid primer sequence (17-mer) 5'-dGTAAAACGACGGCCAGT (SEQ ID NO 31) is synthesized in 0.25 μmole scale on a DNA synthesizer and one phosphorothioate group introduced after the final synthesis cycle (G to T coupling). Sulfurization, deprotection and purification followed standard protocols. Yield was 31.4 nmole (12.6% overall yield), corresponding to 31.4 nmole phosphorothioate groups. Alkylation was performed by dissolving the residue in 31.4 μl TE buffer (0.01 M Tris-HCl, pH 8.0, 0.001 M EDTA) and by adding 16 μl of a solution of 20 mM solution of 2-iodoethanol (320 nmole; 10-fold excess with respect to phosphorothioate diesters) in N,N-dimethylformamide (DMF). The alkylated oligonucleotide was purified by standard reversed phase HPLC (RP-18 Ultraphere, Beckman; column: 4.5 x 250 mm; 100 mM triethyl ammonium acetate, pH 7.0 and a gradient of 5 to 40% acetonitrile).

In a variation of this procedure, the nucleic acid primer containing one or more phosphorothioate phosphodiester bond is used in the Sanger sequencing reactions. The primer-extension products of the four sequencing reactions are purified, cleaved off the solid support, lyophilized and dissolved in 4 μl each of TE buffer pH 8.0 and alkylated by addition of 2 μl of a 20 mM solution of 2-iodoethanol in DMF. It is then analyzed by ES and/or MALDI mass spectrometry.

In an analogous way, employing instead of 2-iodoethanol, *e.g.*, 3-iodopropanol, 4-iodobutanol mass-modified nucleic acid primer are obtained with a mass difference of 14.03, 28.06 and 42.03 daltons respectively compared to the unmodified phosphorothioate phosphodiester-containing oligonucleotide.

EXAMPLE 18: MASS MODIFICATION OF NUCLEOTIDE TRIPHOSPHATES

Mass modification of nucleotide triphosphates at the 2′ and 3′ amino function

Starting material was 2′-azido-2′-deoxyuridine prepared according to literature (Verheyden et al., J. Org. Chem. 36:250, 1971), which was 4,4-dimethoxytritylated at 5′-OH with 4,4-dimethoxytrityl chloride in pyridine and acetylated at 3′-OH with acetic anhydride in a one-pot reaction using standard reaction conditions. With 191 mg (0.71 mmole) 2′-azido-2′-deoxyuridine as starting material, 396 mg (0.65 mmol, 90.8%) 5′-O-(4,4-dimethoxytrityl)-3′-O-acetyl-2′-azido-2′-deoxyuridine was obtained after purification via silica gel chromatography. Reduction of the azido group was performed (Barta et al., Tetrahedron 46:587-94,1990). Yield of 5′-O-(4,4-dimethoxytrityl) 3′-O-acetyl-2′-amino-2′-deoxyuridine after silica gel chromatography was 288 mg (0.49 mmole; 76%). This protected 2′-amino-2′-deoxyuridine derivative (588 mg; 1.0 mmole) was reacted with 2 equivalents (927 mg; 2.0 mmole) N-Fmoc-glycine pentafluorophenyl ester in 10 ml dry DMF overnight at room temperature in the presence of 1.0 mmole (174 μl) N,N-diisopropylethylamine. Solvents were removed by evaporation *in vacuo* and the residue purified by silica gel chromatography. Yield was 711 mg (0.71 mmole; 82%). Detritylation was achieved by a one hour treatment with 80% aqueous acetic acid at room temperature. The residue was evaporated to dryness, co-evaporated twice with toluene, suspended in 1 ml dry acetonitrile and 5′-phosphorylated with $POCl_3$ and directly transformed in a one-pot reaction to the 5′-triphosphate using 3 ml of a 0.5 M solution (1.5 mmole) tetra (tri-n-butylammonium) pyrophosphate in

DMF according to literature. The Fmoc and the 3'-O-acetyl groups were removed by a one-hour treatment with concentrated aqueous ammonia at room temperature and the reaction mixture evaporated and lyophilized. Purification also followed standard procedures by using anion-exchange chromatography on DEAE Sephadex with a linear gradient of triethylammonium bicarbonate (0.1 M - 1.0 M). Triphosphate containing fractions, checked by thin layer chromatography on polyethyleneimine cellulose plates, were collected, evaporated and lyophilized. Yield by UV-absorbance of the uracil moiety was 68% or 0.48 mmole.

A glycyl-glycine modified 2'-amino-2'-deoxyuridine-5'-triphosphate was obtained by removing the Fmoc group from 5'-O-(4,4-dimethoxytrityl)-3'-O-acetyl-2'-N(N-9-fluorenylmethyloxycarbonyl-glycyl)-2'-amino-2'-deoxyuridine by a one-hour treatment with a 20% solution of piperidine in DMF at room temperature, evaporation of solvents, two-fold co-evaporation with toluene and subsequent condensation with N-Fmoc-glycine pentafluorophenyl ester. Starting with 1.0 mmole of the 2'-N-glycyl-2'-amino-2'-deoxyuridine derivative and following the procedure described above, 0.72 mmole (72%) of the corresponding 2'-(N-glycyl-glycyl)-2'-amino-2'-deoxyuridine-5 triphosphate was obtained.

Starting with 5'-O-(4,4-dimethoxytrityl)-3'-O-acetyl-2'-amino-2' deoxyuridine and coupling with N-Fmoc-β-alanine pentafluorophenyl ester, the corresponding 2'-(N-β-alanyl)-2'-amino-2'-deoxyuridine-5' triphosphate are synthesized. These modified nucleoside triphosphates are incorporated during the Sanger DNA sequencing process in the primer extension products. The mass difference between the glycine, β-alanine and glycyl-glycine mass modified nucleosides is, per nucleotide incorporated, 58.06, 72.09 and 115.1 daltons, respectively.

When starting with 5'-O-(4,4-dimethoxytrityl)-3'-amino-2',3'1-dideoxythymidine, the corresponding 3'-(N-glycyl)-3'-amino-, 3'-(-N-glycyl-glycyl)-3'-amino-, and 3'-(N-β-3-alanyl)-3'-amino-2',3'-dideoxythymidine-5'-triphosphates can be obtained. These mass-

modified nucleoside triphosphates serve as a terminating nucleotide unit in the Sanger DNA sequencing reactions providing a mass difference per terminated fragment of 58.06, 72.09 and 115.1 daltons respectively when used in the multiplexing sequencing mode. The mass differentiated fragments are analyzed by ES and/or MALDI mass spectrometry.

Mass modification of nucleotide triphosphates at C-5 of the heterocyclic base: 0.281 g (1.0 mmole) 5-(3-Aminopropynyl-I)-2'-deoxyuridine was reacted with either 0.927 g (2.0 mmole) N-Fmoc-glycine pentafluorophenylester or 0.955g (2.0 mmole) N-Fmoc-β-alanine pentafluorophenyl ester in 5 ml dry DMF in the presence of 0.129 g N, N-diisopropylethylamine (174 μl; 1.0 mmole) overnight at room temperature. Solvents were removed by evaporation *in vacuo* and the condensation products purified by flash chromatography on silica gel (Still et al., J. Org., Chem. 43: 2923-25, 1978). Yields were 476 mg (0.85 mmole; 850%) for the glycine and 436 mg (0.76 mmole; 76%) for the β-alanine derivatives. For the synthesis of the glycyl-glycine derivative, the Fmoc group of 1.0 mmole Fmoc-glycine-deoxyuridine derivative was removed by one-hour treatment with 20% piperidine in DMF at room temperature. Solvents were removed by evaporation in vacuo, the residue was coevaporated twice with toluene and condensed with 0.927 g (2.0 mmole) N-Fmoc-glycine pentafluorophenyl ester and purified as described above. Yield was 445 mg (0.72 mmole; 72%). The glycyl-, glycyl-glycyl- and β-alanyl-2-deoxyuridine derivatives, N-protected with the Fmoc group were transformed to the 3'-O-acetyl derivatives by tritylation with 4,4-dimethoxytrityl chloride in pyridine and acetylation with acetic anhydride in pyridine in a one-pot reaction and subsequently detritylated by one-hour treatment with 80% aqueous acetic acid according to standard procedures. Solvents were removed, the residues dissolved in 100 ml chloroform and extracted twice with 50 ml 10% sodium bicarbonate and once with 50 ml water, dried with sodium sulfate, the solvent evaporated and the residues purified by flash chromatography on silica gel. Yields

were 361 mg (0.60 mmole; 71%) for the glycyl-, 351 mg (0.57 mmole; 75%) for the β-alanyl- and 323 mg (0.49 mmole; 68%) for the glycyl-glycyl-3-0'-acetyl-2'-deoxyuridine derivatives, respectively. Phosphorylation at the 5'-OH with $POCl_3$, transformation into the 5'-triphosphate by *in situ* reaction with tetra (tri-n-butylammonium) pyrophosphate in DMF, 3'-de-0-acetylation, cleavage of the Fmoc group, and final purification by anion-exchange chromatography on DEAE-Sephadex was performed and yields according to UV-absorbance of the uracil moiety were 0.41 mmole 5-(3-(N-glycyl)-amidopropynyl-l)-2'-deoxyuridine-5'-triphosphate (84%), 0.43 mmole 5-(3-(N-β-alanyl)-amidopropynyl-l)-2'-deoxyuridine-5'-triphosphate (75%) and 0.38 mmole 5-(3-(N-glycyl-glycyl)-amidopropynyl-l)-2'-deoxyuridine-5'-triphosphate (78%). These mass-modified nucleoside triphosphates were incorporated during the Sanger DNA sequencing primer-extension reactions.

When using 5-(3-aminopropynyl)-2',3'-dideoxyuridine as starting material and following an analogous reaction sequence the corresponding glycyl-, glycyl-glycyl and β-alanyl-2',-3'-dideoxyuridine-5'-triphosphates were obtained in yields of 69%, 63% and 71%, respectively. These mass-modified nucleoside triphosphates serve as chain-terminating nucleotides during the Sanger DNA sequencing reactions. The mass-modified sequencing ladders are analyzed by either ES or MALDI mass spectrometry.

Mass modification of nucleotide triphosphates: 727 mg (1.0 mmole) of N^6-(4-tert-butylphenoxyacetyl)-8-glycyl-5'-(4,4-dimethoxytrityl)-2'-deoxyadenosine or 800 mg (1.0 mmole) N^6-4-tert-butylphenoxyacetyl)-8-glycyl-glycyl-5'-(4,4-dimethoxytrityl)-2'-deoxyadenosine prepared according to literature (Köster et al., Tetrahedron 37:362, 1981) were acetylated with acetic anhydride in pyridine at the 3'-OH, detritylated at the 5'-position with 80% acetic acid in a one-pot reaction and transformed into the 5'-triphosphates via phosphorylation with $POCl_3$ and reaction *in situ* with tetra(tri-n-butylammonium) pyrophosphate.

Deprotection of the N^6-tert-butylphenoxyacetyl, the 3'-O-acetyl and the O-methyl group at the glycine residues was achieved with concentrated aqueous ammonia for ninety minutes at room temperature. Ammonia was removed by lyophilization and the residue washed with dichloromethane, solvent removed by evaporation *in vacuo* and the remaining solid material purified by anion exchange chromatography on DEAE-Sephadex using a linear gradient of triethylammonium bicarbonate from 0.1 to 1.0 M. The nucleoside triphosphate containing fractions (checked by TLC on polyethyleneimine cellulose plates) were combined and lyophilized. Yield of the 8-glycyl-2'-deoxyadenosine-5'-triphosphate (determined by UV-absorbance of the adenine moiety) was 57% (0.57 mmole). The yield for the 8-glycyl-glycyl-2'-deoxyadenosine-5'-triphosphate was 51% (0.51 mmole). These mass-modified nucleoside triphosphates were incorporated during primer-extension in the Sanger DNA sequencing reactions.

When using the corresponding N6-(4-tert-butylphenoxyacetyl)-8-glycyl- or -glycyl-glycyl-5'-O-(4,4-dimethoxytrityl)-2',3'-dideoxyadenosine derivatives as starting materials (for the introduction of the 2',3'-function: Seela et al., Helvetica Chimica Acta 74:1048-58, 1991). Using an analogous reaction sequence, the chain-terminating mass-modified nucleoside triphosphates 8-glycyl- and 8-glycyl-glycyl-2'3'-dideoxyadenosine-5'-triphosphates were obtained in 53 and 47% yields, respectively. The mass-modified sequencing fragment ladders are analyzed by either ES or MALDI mass spectrometry.

EXAMPLE 19: MASS MODIFICATION OF NUCLEOTIDES BY ALKYLATION AFTER SANGER SEQUENCING

2',3'-dideoxythymidine-5'-(alpha-S)-triphosphate was prepared according to published procedures (for the alpha-S-triphosphate moiety: Eckstein et al., Biochemistry 15:1685, 1976 and Accounts Chem. Res. 12:204, 1978; and for the 2',-3'-dideoxy moiety: Seela et al., Helvetica Chimica Acta 74:1048-58, 1991). Sanger DNA sequencing reactions

employing 2'-deoxythymidine-5'-(alpha-S)-triphosphate are performed according to standard protocols. When using 2',3'-dideoxythymidine-5'-(alpha-S)-triphosphates, this is used instead of the unmodified 2',3'-dideoxythymidine-5'-triphosphate in standard Sanger DNA sequencing. The template (2 pmole) and the nucleic acid M13 sequencing primer (4 pmole) are annealed by heating to 65°C in 100 μl of 10 mM Tris-HCl pH 7.5, 10 mM MgCl2, 50 mM NaCl, 7 mM dithiothreitol (DTT) for 5 minutes and slowly brought to 37°C during a one hour period. The sequencing reaction mixtures contain, as exemplified for the T-specific termination reaction, in a final volume of 150 μl, 200 μM (final concentration) each of dATP, dCTP, dTTP, 300 μM c7-deaza-dGTP, 5 μM 2',3'dideoxythymidine-5'-(alpha-S)-triphosphate and 40 units Sequenase. Polymerization is performed for 10 minutes at 37°C, the reaction mixture heated to 70°C to inactivate the Sequenase, ethanol precipitated and coupled to thiolated Sequelon membrane disks (8 mm diameter). Alkylation is performed by treating the disks with 10 μl of 10 mM solution of either 2-iodoethanol or 3-iodopropanol in NMM (N-methylmorpholine/water/2-propanol, 2/49/49, v/v/v) (three times), washing with 10 μl NMM (three times) and cleaving the alkylated T-terminated primer-extension products off the support by treatment with DTT. Analysis of the mass-modified fragment families is performed with either ES or MALDI mass spectrometry.

EXAMPLE 20: MASS MODIFICATION OF AN OLIGONUCLEOTIDE

This method, in addition to mass modification, also modifies the phosphate backbone of the nucleic acids to a non-ionic polar form. Oligonucleotides can be obtained by chemical synthesis or by enzymatic synthesis using DNA polymerases and α-thio nucleoside triphosphates.

This reaction was performed using DMT-TpT as a starting material but the use of an oligonucleotide with an alpha thio group is also appropriate. For thiolation, 45 mg (0.05 mM) of compound 1 (Figure 15), is dissolved in 0.5 ml acetonitrile and thiolated in a 1.5 ml tube with 1,1-diozo-1-H-benzo[1,2]dithio-3-on (Beaucage reagent). The reaction was

allowed to proceed for 10 minutes and the produce is concentrated by thin layer chromatography with the solvent system dichloromethane/96% ethanol/pyridine (87%/13%/1% v/v/v). The thiolated compound 2 (Figure 15) is deprotected by treatment with a mixture of concentrated aqueous ammonia/acetonitrile (1/1; v/v) at room temperature. This reaction is monitored by thin layer chromatography and the quantitative removal of the beta cyanoethyl group was accomplished in one hour. This reaction mixture was evaporated *in vacuo*.

To synthesize the S-(2-amino-2-oxyethyl)thiophosphate triester of DMT-TpT (compound 4), the foam obtained after evaporation of the reaction mixture (compound 3) was dissolved in 0.3 ml acetonitrile/pyridine (5/1; v/v) and a 1.5 molar excess of iodoacetamide added. The reaction was complete in 10 minutes and the precipitated salts were removed by centrifugation. The supernatant is lyophilized, dissolved in 0.3 ml acetonitrile and purified by preparative thin layer chromatography with a solution of dichloromethane/96% ethanol (85%/15%; v/v). Two fractions are obtained which contain one of the two diastereoisomers. The two forms were separated by HPLC.

EXAMPLE 21: MALDI-MS ANALYSIS OF A MASS-MODIFIED OLIGONUCLEOTIDE.

A 17-mer was mass-modified at C-5 of one or two deoxyuridine moieties. 5 [13-(2-Methoxyethoxyl)-tridecyne-1-yl]-5'-O-(4,4'-dimethoxytrityl)-2'-deoxyuridine-3'-β-cyanoethyl-N,N-diisopropylphosphoamidite was used to synthesize the modified 17-mers.

The modified 17-mers were:

where X = $-C\equiv C-(CH_2)_{11}-OH$
(unmodified 17-mer: molecular mass: 5273)

The samples were prepared and 500 fmol of each modified 17-mer was analyzed using MALDI-MS. Conditions used were reflectron positive ion mode with an acceleration of 5 kV and postacceleration of 20 kV. The MALDI-TOF spectra which were generated were superimposed and are shown in Figure 16. Thus, mass modification provides a distinction detectable by mass spectrometry which can be used to identify base sequence information.

Other embodiments and uses of the invention will be apparent to those skilled in the art from consideration of the specification and practice of the invention disclosed herein. The specification and examples should be considered exemplary only with the true scope and spirit of the invention indicated by the following claims.

09345409.061402